Accretion Driven Evolution of Quasars and Black Holes: Theoretical Models

Adam Steed and David H. Weinberg

Department of Astronomy, The Ohio State University, Columbus, OH 43210

asteed,dhw@astronomy.ohio-state.edu

ABSTRACT

We present a flexible framework for constructing physical models of quasar evolution that can incorporate a wide variety of observational constraints, such as multi-wavelength luminosity functions, estimated masses and accretion rates of active black holes, space densities of quasar host galaxies, clustering measurements, and the mass function of black holes in the local universe. The central actor in this formulation is the accretion rate distribution $p(\dot{m}|M,z)$, the probability that a black hole of mass M at redshift z accretes at a rate $\dot{m}$ in Eddington units. Given a model of accretion physics that specifies the radiative efficiency and SED shape as a function of $\dot{m}$, the quasar luminosity function (QLF) is determined by a convolution of $p(\dot{m}|M,z)$ with the black hole mass function $n(M,z)$. In the absence of mergers, $p(\dot{m}|M,z)$ also determines the full evolution of $n(M,z)$, given a "boundary value" of $n(M)$ at some redshift. If $p(\dot{m}|z)$ is independent of mass, then the asymptotic slopes of the QLF match the asymptotic slopes of $n(M)$, and $n(M)$ evolves in a self-similar fashion, retaining its shape while shifting to higher masses. Matching the observed decline of the QLF "break" luminosity at $z < 2$ requires either a shift in $p(\dot{m}|z)$ that increases the relative probability of low accretion rates or an evolving mass dependence of $p(\dot{m}|M,z)$ that preferentially shuts off accretion onto high mass black holes at low z. These two scenarios make different predictions for the masses and accretion rates of active black holes. If the first mechanism dominates, then the QLF changes character between $z = 2$ and $z = 0$, shifting from a sequence of black hole mass towards a sequence of $L/L_{\rm Edd}$. We use our framework to compare the predictions of five models that illustrate different assumptions about the quasar population: two dominated by unobscured thin-disk accretion with short and long quasar lifetimes, respectively, one with a 4:1 ratio of obscured to unobscured systems, one with substantial black hole merger activity at low redshift, and one with substantial low redshift growth in radiatively inefficient flows. We discuss the observational advances that would be most valuable for distinguishing such models and for pinning down the physics that drives black hole and quasar evolution.

Subject headings: quasars: general

Submitted to The Astrophysical Journal, February 2, 2008.

1. Introduction

The study of the quasar and AGN populations has been transformed in recent years by ambitious new optical (e.g., Boyle et al. 2000; Schneider et al. 2002; Wolf et al. 2003), X-ray (e.g., Brandt et al. 2001; Giacconi et al. 2002; Anderson et al. 2003; Ueda et al. 2003), and radio (e.g., White et al. 2000) surveys, by the recognition that low efficiency accretion modes may become important when the accretion rate itself is low (e.g., Narayan et al. 1998 and references therein), by detailed studies of low luminosity AGN in the local universe (e.g., Ho 2001), and, perhaps most of all, by the accumulating evidence that supermassive black holes are ubiquitous in the bulges of present-day galaxies (e.g., Richstone et al. 1998; Merritt & Ferrarese 2000; Gebhardt et al. 2000). The dynamical studies of nearby galaxies strengthen the long-standing hypothesis that quasars are powered by black hole accretion (e.g., Lynden-Bell 1969; Rees 1978), and the "demography" of the local black hole population provides a powerful constraint on models of quasar evolution and its connection to galaxy evolution. These developments have inspired increasingly sophisticated theoretical models that place quasar evolution in the context of hierarchical clustering models for the formation of dark matter halos and galaxies (e.g., Kauffmann & Haehnelt 2000; Cavaliere & Vittorini 2002; Haiman & Loeb 2001; Wyithe & Loeb 2003).

This paper presents a physically motivated calculational framework that is intermediate in complexity between such *ab initio* models of the quasar population and older descriptions in terms of "luminosity evolution" or "density evolution." The central actor in our formulation is the accretion probability distribution $p(\dot{m}|M, z)$, the probability that a black hole of mass M at redshift z is accreting mass at a rate $\dot{m}$ in Eddington units (discussed in §2 below). The key supporting players are the black hole mass function $n(M, z)$ and a physical model of accretion that predicts the radiative efficiency for a given $\dot{m}$.[1] An example of an accretion model would be thin-disk accretion with efficiency $\epsilon \equiv L/\dot{M}c^2 \sim 0.1$ when $\dot{m} \sim 1$, changing to low efficiency advection-dominated (ADAF) accretion when $\dot{m}$ is below some critical value. At a given redshift, $p(\dot{m}|M)$ and $n(M)$ together determine the quasar luminosity function, and $p(\dot{m}|M)$ also determines the accretion driven growth of the black hole population, and hence the evolution of $n(M)$. Thus, given physical assumptions about radiative efficiencies and a "boundary condition" specifying $n(M)$ at one redshift, the history of $p(\dot{m}|M)$ determines the complete evolution of the black hole population and the quasar luminosity function. An essential caveat is that mergers of black holes following mergers of their host galaxies could alter $n(M)$ independently of the accretion characterized by $p(\dot{m}|M)$.

The simplest scenario connecting black hole and quasar evolution is that black holes "shine as they grow": a luminous quasar is powered by a black hole radiating at near-Eddington luminosity with efficiency $\epsilon \sim 0.1$, and no significant growth occurs in a non-luminous phase. In this case, the bolometric luminosity function at a given redshift is just $\Phi(L) = f_{\rm on} n(L/l) l^{-1}$, where l is the

[1] Henceforth, we will usually drop the explicit dependence on z and refer only to $p(\dot{m}|M)$ or $n(M)$, but these should always be understood to refer to the distribution at some particular redshift.

(universal) ratio of Eddington luminosity to black hole mass and $f_{\rm on}$ is the fraction of black holes that are accreting. In a time interval Δt, black holes on average increase their mass by a factor $\exp(f_{\rm on}\Delta t/t_g)$, where $t_g = M/(L_{\rm Edd}/\epsilon c^2) = 4.5 \times 10^7(\epsilon/0.1)$ yr is the e-folding time for growth at the Eddington luminosity (Salpeter 1964; see discussion in §2). The mass density in black holes at the present day is simply related to the emissivity of the quasar population integrated over luminosity and redshift, $\rho_{\rm BH} = \int_0^{t_0} U(t)dt/\epsilon c^2$ (Soltan 1982, updated by, e.g., Chokshi & Turner 1992; Richstone et al. 1998; Yu & Tremaine 2002; Fabian 2003). In our language, this is a model in which all active black holes have the same radiative efficiency and $p(\dot{m}|M) = f_{\rm on}\delta_D(\dot{m} - 1)$, where δ_D is the Dirac-delta function. The evolution of quasars and black holes is determined by a boundary condition on $n(M)$ and the redshift history of the active fraction $f_{\rm on}(z)$. The "quasar era" $z \sim 2 - 4$ when the emissivity of the population peaks is also the era in which today's black holes grew to their current mass.

This simple scenario may not be too far from the truth, but the possible complications raise a number of questions. Did today's black holes gain a significant fraction of their mass through low-$\dot{m}$, low efficiency, ADAF-type accretion, thus growing at low luminosity? Have black hole mergers substantially altered $n(M)$, leaving the integrated density $\rho_{\rm BH}$ fixed but changing the relative numbers of high and low mass black holes? Are some quasars accreting mass at super-Eddington rates, radiating at high luminosity but low efficiency in "smothered," optically thick ADAF modes (Katz 1977; Begelman 1978; Abramowicz et al. 1988)? Do some quasars radiate substantially above the Eddington limit (Begelman 2002)? Is a significant fraction of quasar activity obscured by gas and dust, as hypothesized in synthesis models of the X-ray background (e.g., Setti & Woltjer 1989; Comastri et al. 1995; Fiore et al. 1999; Fabian & Iwasawa 1999; Gilli et al. 2001), thus redistributing bolometric luminosity from the optical-UV-soft X-ray to the far-IR? Are low luminosity AGNs powered mainly by low mass black holes radiating at Eddington luminosity, by more massive black holes with thin-disk efficiencies but sub-Eddington accretion rates, or by still more massive black holes with sub-Eddington accretion rates *and* low efficiency? The methods developed here provide useful tools for addressing these questions, allowing us to construct concrete, quantitative models that answer them in different ways, then examine how observational data might distinguish among such models.

Our framework complements, but by no means replaces, the *ab initio* approach that connects the evolution of quasars and black holes to that of the underlying dark halo and galaxy populations. This approach has yielded many valuable insights, including the recognition that the rise of the quasar population probably traces the formation of the first dark halos large enough to host massive black holes, that the rapid decline of the population at low redshift probably reflects the combined impact of declining galaxy interaction rates and decreasing gas supplies in quasar hosts, and that the clustering of quasars with themselves or with galaxies can provide a valuable diagnostic of typical quasar lifetimes (e.g., Efstathiou & Rees 1988; Haehnelt & Rees 1993; Haehnelt et al. 1998; Salucci et al. 1999; Kauffmann & Haehnelt 2000; Cavaliere & Vittorini 2000; Haiman & Hui 2001; Martini & Weinberg 2001; Menou et al. 2001; Wyithe & Loeb 2003). In combination with semi-analytic models of galaxy formation, these quasar evolution models can also predict the properties and environments of quasar hosts and the relation between the properties of present day galaxies and the masses of their central black holes. However, the models necessarily rely on specific assumptions about the mechanisms that trigger quasar activity and the accretion rates that these mechanisms produce. To put things in our terms, the *ab initio* models adopt a particular set of hypotheses about quasar activity in order to predict $p(\dot{m}|M,z)$ from first principles.

Our framework is designed to model observational data in a flexible way with relatively few assumptions, while retaining the basic physical picture of black hole accretion that underlies nearly all modern interpretations of the quasar population. One hope is that measurements of the luminosity function and the local black hole mass function will eventually allow us to integrate backwards in time and determine $p(\dot{m}|M,z)$ empirically, drawing on a variety of observations to test the assumptions that enter such a reconstruction. We may find that the data are not powerful enough to tie down $p(\dot{m}|M,z)$ without some *a priori* constraints on its expected form, but that finding in itself would be a valuable, if disappointing, lesson. More generally, we hope to illuminate the connections between black hole evolution and quasar activity and learn what observations can and cannot tell us about these connections.

The Haehnelt et al. (1998) paper has had the strongest impact on our thinking about these issues, but the most direct antecedent that we know of to the approach taken here is the lucid paper of Small & Blandford (1992). They adopted a similar description of black hole evolution and its connection to quasar activity, and they applied this description to the observational data available at the time. Advances in the observational data and the theoretical models of accretion make this an opportune time to revive and extend this approach. Yu & Tremaine (2002) have recently used a similar method in assessing constraints on black hole accretion and mergers, though their assumptions and goals are more restrictive than ours — in particular, they assume that quasars radiate at Eddington luminosity and thus that $p(\dot{m})$ consists of δ-functions at $\dot{m}=1$ and $\dot{m}=0$.

The difference in the form of $p(\dot{m})$ is one of the most significant differences between the models presented in this paper and most models of the quasar population in the literature. These typically assume that $p(\dot{m})$ for $\dot{m}>0$ is sharply peaked at some value close to Eddington, such as a δ-function (e.g., Small & Blandford 1992) or a spike followed by an exponential decline (e.g., Haehnelt et al. 1998; Kauffmann & Haehnelt 2000). A sharply peaked $p(\dot{m})$ could arise physically if the central black hole typically plays a large role in controlling its own fuel supply, through feedback or influence on stellar dynamics. However, we think it is more likely that fueling is driven by galactic scale events — galaxy mergers, interactions, and bar formation, for example — that are minimally influenced by the central black hole, and therefore do not "know" that they should feed it at any particular rate. In particular, it seems reasonable that for every major event that leads to Eddington-like fueling of a central black hole there are many minor events that fuel it at a sub-Eddington rate. The particular functional forms that we adopt here to represent this scenario, a power-law or broken power-law $p(\dot{m})$ between some $\dot{m}_{\rm min}$ and some $\dot{m}_{\rm max}$, are arbitrary, and chosen largely for mathematical convenience, but they reflect this general thinking about the process of quasar fueling. In the long run, one goal of our effort is to test observationally whether $p(\dot{m})$ is in fact a broad function or a peaked function, which would in turn have implications for the mechanisms of quasar fueling. At a qualitative level, the wide $L/L_{\rm Edd}$ distribution of AGN activity in the local universe (e.g, Ho 2001) seems to support the idea of a broad distribution of accretion rates.

In this paper we will keep our contact with observations relatively loose, focusing instead on presenting our framework in a clear way and illustrating how observations might distinguish among different scenarios for the quasar population. We will carry out a detailed analysis of multi-wavelength measurements of the quasar luminosity function and constraints on active black hole masses and the local black hole mass function a subsequent paper. We adopt a cosmological model with $\Omega_m = 1$ and $h \equiv H_0/100\ {\rm km\,s^{-1}\,Mpc^{-1}} = 0.5$ because all of the observational papers include results for this model, and not always for the low density, Λ-dominated, $h \sim 0.7$ model favored by recent cosmological observations. Cosmology affects our models indirectly through its influence on observationally inferred luminosity functions and directly through the time-redshift relation. We would not expect a change of cosmological model to have any qualitative impact on our results, and we expect that the quantitative impact could be compensated by modest changes to mass scales and accretion rates, especially since the age of the universe is similar for $\Omega_m = 1$, $h = 0.5$, and for $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$, $h = 0.7$.

We present the definitions and key equations of our framework in the following section. We then present mathematical results for luminosity functions in §3 and for luminosity and black hole mass evolution in §4, focusing on analytically solvable cases that illustrate general points. In §5 we construct five illustrative models of the quasar population, each designed to match the observed evolution of the optical luminosity function but differing from one another in the typical quasar lifetime or in the importance of mergers, ADAF growth, or obscuration. We show how measurements of the black hole mass function, luminosity functions at other wavelengths, the masses and accretion rates of active black holes, and the space density of host galaxies might distinguish among these scenarios. We summarize our results in §6. This is a long paper, and a reader who wants to get quickly to the main points can read §2, look through the figures and captions, and read §6. The definitions of the models illustrated in Figures 8–17 are summarized in Table 2 and its accompanying caption.

2. Framework

2.1. Definitions and assumptions

We define l to be the ratio of a black hole's Eddington luminosity to its mass,

$$l \equiv \frac{L_{\rm Edd}}{M} = \frac{4\pi G m_p c}{\sigma_T} = 1.26 \times 10^{38} {\rm erg\ sec^{-1}} {M_\odot}^{-1}. \qquad (1)$$

We often scale the radiative efficiencies to the value 0.1 that is typically adopted for thin-disk accretion,

$$\epsilon \equiv \frac{L}{\dot{M}c^2} = 0.1\epsilon_{0.1}, \tag{2}$$

where $\dot{M}$ is the mass accretion rate and L is the bolometric luminosity. We define the Eddington accretion rate to be the mass accretion rate for which a black hole *with radiative efficiency* $\epsilon_{0.1} = 1$ has Eddington luminosity,

$$\dot{M}_{\rm Edd} \equiv \frac{L_{\rm Edd}}{0.1c^2} \approx 22\left(\frac{M}{10^9 M_\odot}\right) M_\odot\ {\rm yr}^{-1}, \tag{3}$$

and the dimensionless accretion rate

$$\dot{m} \equiv \frac{\dot{M}}{\dot{M}_{\rm Edd}}\ . \tag{4}$$

This definition of $\dot{M}_{\rm Edd}$ in terms of a fixed, thin-disk efficiency is common in the literature, but not universal. A black hole accreting at $\dot{M}_{\rm Edd}$ grows in mass exponentially, with an e-folding timescale

$$t_s \equiv \frac{M}{\dot{M}_{\rm Edd}} = 4.5 \times 10^7 {\rm yr}. \tag{5}$$

For $\epsilon_{0.1} = 1$, t_s is equal equal to the Salpeter (1964) timescale for growth at the Eddington luminosity; note, however, that we define t_s to be 4.5×10^7 yr independent of efficiency and of $L/L_{\rm Edd}$. With these definitions, a black hole of mass M accreting at a dimensionless rate $\dot{m}$ with radiative efficiency $\epsilon = 0.1\epsilon_{0.1}$ has bolometric luminosity

$$L = 0.1\epsilon_{0.1}\dot{M}c^2 = \epsilon_{0.1}\dot{m}lM\ . \tag{6}$$

At each redshift, we define the black hole mass function $n(M)$ such that $n(M)dM$ is the comoving space density of black holes in the mass range $M \rightarrow M + dM$; the units of $n(M)$ are thus number per comoving volume per unit mass. In our plots, we usually show $Mn(M)$, the comoving space density (with units Mpc^{-3}) in an interval $d\ln M$, rather than $n(M)$ itself. We define the accretion probability distribution $p(\dot{m}|M)$ such that $p(\dot{m}|M)d\dot{m}$ is the probability that a black hole of mass M has an accretion rate in the range $\dot{m} \rightarrow \dot{m} + d\dot{m}$. In our subsequent calculations, we will frequently consider the restricted and analytically convenient class of models in which $p(\dot{m}|M) = p(\dot{m})$, i.e., with accretion probability distribution independent of mass. Since more massive black holes reside in more massive galaxies that have larger internal gas supplies and can cannibalize larger companions, this assumption could be a reasonable approximation to the real universe (see further discussion in §4.2.2 below). However, it is at best a convenient approximation, and we will try to be mindful of its limitations.

Many theoretical models of quasar evolution specify $p(\dot{m})$ implicitly through a typical "light curve" $L(t)$ that follows each triggering event. The probability $p(\dot{m})d\dot{m}$ can be identified with the fraction of time that $L/L_{\rm Edd}$ is in the range $\epsilon_{0.1}\dot{m} \rightarrow \epsilon_{0.1}(\dot{m} + d\dot{m})$. However, the relation between light curves and $p(\dot{m})$ distributions is many-to-one, even for constant efficiency. For example, if every quasar lights up at the Eddington luminosity and declines exponentially thereafter, with $\epsilon_{0.1} = 1$ throughout, then there is constant accretion probability per logarithmic interval of $\dot{m}$, and thus $p(\dot{m}) \propto \dot{m}^{-1}$ for $\dot{m} \leq 1$. But the same $p(\dot{m})$ could correspond to an "on-off" activity model where the luminosity of each quasar is constant for a fixed time while the number of quasars per luminosity interval is proportional to $(L/L_{\rm Edd})^{-1}$.

The model of accretion physics specifies the probability that a black hole of mass M and accretion rate $\dot{m}$ has bolometric radiative efficiency $\epsilon_{0.1}$. Since most of the properties of accretion flows scale in a fairly simple way with mass, it is reasonable to expect that this probability distribution is independent of M. To simplify our calculations, we will also assume that all black holes of the same $\dot{m}$ have the same $\epsilon_{0.1}$, i.e., we will assume that the probability distribution of $\epsilon_{0.1}$ has zero width at given $\dot{m}$. Outside of a small range of $\dot{m}$ where black holes might cycle between thin-disk accretion and an ADAF-type flow, this assumption seems plausible, though one could imagine that a range of black hole spins could induce a range of $\epsilon_{0.1}$ values even at fixed $\dot{m}$. We will generally assume that $\epsilon_{0.1} = 1$ in a range $\dot{m}_{\rm crit} \leq \dot{m} \leq 1$, where $\dot{m}_{\rm crit}$ is a critical value of the accretion rate below which thin-disk accretion gives way to some lower efficiency mode. For stellar mass black holes, which exhibit transitions among accretion modes, $\dot{m}_{\rm crit}$ may be as high as 0.09 (Esin et al. 1997), but for supermassive black holes the value is more uncertain and probably lower (R. Narayan, private communication). We will adopt $\dot{m}_{\rm crit} = 0.01$. Below this threshold, we will assume $\epsilon_{0.1} = (\dot{m}/\dot{m}_{\rm crit})$, a scaling motivated by ADAF models (Narayan et al. 1998).

If the accretion rate is determined by large scale gas flows beyond the influence of the black hole itself, then there is no particular reason that $\dot{m}$ should not exceed one. In such situations, we will assume that the accretion proceeds in a lower efficiency, "smothered" mode (Katz 1977; Begelman 1978; Abramowicz et al. 1988), so that the black hole radiates at the Eddington luminosity — in other words, $\epsilon_{0.1} = \dot{m}^{-1}$ when the accretion rate is super-Eddington. It is not clear that these smothered accretion modes are stable enough to exist in nature, and it may be that black holes in this situation drive the excess gas out of the nucleus altogether rather than accepting mass at a super-Eddington rate, thus regulating the accretion rate so that $\dot{m}$ never exceeds one. The distinction between these two pictures is usually not important for our purposes, provided that the black holes radiate at the Eddington luminosity in either case, but it does make a difference if super-Eddington accretion makes a significant contribution to black hole mass growth. It has also been suggested (Begelman 2002) that black holes can radiate substantially above the Eddington luminosity (an order of magnitude or more) — this *would* make a difference to our predictions, as we discuss briefly in §6.

To summarize, we generally assume that the bolometric radiative efficiency is

$$\epsilon_{0.1}(\dot{m}) = \begin{cases} (\dot{m}/\dot{m}_{\rm crit}) & \text{for } \dot{m} < \dot{m}_{\rm crit} \\ 1 & \text{for } \dot{m}_{\rm crit} \leq \dot{m} \leq 1 \\ \dot{m}^{-1} & \text{for } \dot{m} > 1, \end{cases} \qquad (7)$$

with $\dot{m}_{\rm crit} = 0.01$. We will also usually assume that $p(\dot{m})$ is non-zero only over some range $\dot{m}_{\rm min}$ to $\dot{m}_{\rm max}$, and in some simplified cases we will choose this range so that only the thin-disk regime with $\epsilon_{0.1}(\dot{m}) = 1$ contributes. To calculate luminosities in particular wavebands, we will incorporate luminosity fractions F_ν, which in some cases we allow to depend on $\dot{m}$ or to vary from one black hole to another. We discuss our assumptions about F_ν as they arise.

2.2. The luminosity function and the black hole evolution equation

The QLF is obtained from the black hole mass function and the accretion probability distribution by a straightforward counting argument. Black holes in the mass range $M \rightarrow M + dM$ with accretion rate $\dot{m}$ correspond to quasars with luminosity in the range $L \rightarrow L + dL$ with $L = \epsilon_{0.1}\dot{m}lM$ and $dL = (\epsilon_{0.1}\dot{m}l)dM$ (eq. 6). The comoving space density of black holes in this mass range is $n(M)dM$, and the corresponding density of quasars with luminosity $L \rightarrow L + dL$, is obtained by multiplying by the accretion probability $p(\dot{m}|M)$ and integrating over $\dot{m}$:

$$\Phi(L)dL = \int_{\dot{m}_{\rm min}}^{\dot{m}_{\rm max}} d\dot{m}\, p(\dot{m}|M)n(M)\frac{dL}{\epsilon_{0.1}\dot{m}l}, \qquad M = \frac{L}{\epsilon_{0.1}\dot{m}l}, \tag{8}$$

where the integral covers the full range over which $p(\dot{m}|M)$ is non-zero. One can obtain a mathematically equivalent expression by identifying the luminosity range dL with the range of accretion rates $d\dot{m} = dL/(\epsilon_{0.1}lM)$ at fixed M, then integrating over masses. The above argument is slightly complicated by the possibility that $\epsilon_{0.1}$ depends on $\dot{m}$, but provided that $\epsilon_{0.1}\dot{m}$ is a monotonic function of $\dot{m}$, the probability density transformation $p(\dot{m})d\dot{m} = p(\epsilon_{0.1}\dot{m})d(\epsilon_{0.1}\dot{m})$ still leads to equation (8). For our standard assumption about super-Eddington accretion, $\epsilon_{0.1}\dot{m} = 1$ for $\dot{m} > 1$, the luminosity function of super-Eddington accretors is

$$\Phi_{\rm SE}(L)dL = n(L/l)\frac{dL}{l}\int_1^{\dot{m}_{\rm max}} d\dot{m}\, p(\dot{m}|M = L/l), \tag{9}$$

so equation (8) remains valid when $\dot{m}_{\rm max} > 1$. Allowing a range of efficiencies at fixed $\dot{m}$ would broaden the luminosity function, since one would then convolve $n(M)$ with $p(\epsilon_{0.1}\dot{m})$ instead of $p(\dot{m})$ to obtain $\Phi(L)$.

Equation (8) gives the bolometric luminosity function in terms of $n(M)$ and $p(\dot{m}|M)$ at a particular redshift. If accretion is the only process contributing to black hole growth, then the evolution of $n(M)$ is determined by a simple continuity equation,

$$\frac{\partial n(M,t)}{\partial t} = -\frac{\partial (n\langle \dot{M}(t)\rangle)}{\partial M} = -\frac{1}{t_s}\frac{\partial (nM\langle \dot{m}(M,t)\rangle)}{\partial M}, \tag{10}$$

which follows from considering the rate at which black holes are leaving and entering a mass range $M \rightarrow M + dM$ (Small & Blandford 1992). (The last equality follows from equations [4] and [5], which imply that $\dot{M} = \dot{m}M/t_s$.) The important simplification that follows from the continuity argument is that the evolution of $n(M)$ depends on $p(\dot{m}|M)$ only through the mean accretion rate, $\langle\dot{m}(M,t)\rangle = \int d\dot{m}\, \dot{m}\, p(\dot{m}|M,t)$. In any given time interval, some black holes will grow faster than average and some will grow slower, but the mass function is an average over the full population, and its evolution depends only on the average rate at which black holes move from one mass range to another.

If $\langle\dot{m}(t)\rangle$ is independent of M, then it can be moved out of the derivative on the r.h.s. of equation (10), and in this case the solution to the evolution equation is remarkably simple:

$$n(M,t) = F\left(\frac{M}{M_*}\right)\frac{M_{*,i}}{M_*}, \qquad M_*(t) = M_{*,i}\exp\left(\int_{t_i}^{t}\langle\dot{m}(t)\rangle\frac{dt}{t_s}\right), \tag{11}$$

where $F(x)$ is an arbitrary function and M_* is any fiducial scale in the mass function. This solution can be verified by direct substitution into equation (10), but its physical basis is easy to see. Since $\langle\dot{m}(t)\rangle$ is independent of mass, and only $\langle\dot{m}\rangle$ matters rather than the full distribution $p(\dot{m})$, the evolution is equivalent to that of a population in which all black holes accrete mass at the rate $\dot{M} = \langle\dot{m}(t)\rangle\dot{M}_{\rm Edd} = \langle\dot{m}(t)\rangle M/t_s$. In this case, every black hole grows by the exponential factor on the r.h.s. of equation (11), so the scale of the mass function simply shifts; the normalization drops in proportion to $1/M_*$ because the width of the differential mass range dM occupied by a given set of black holes increases as the black hole masses themselves increase. The simplicity of the solution (11) makes models in which $p(\dot{m})$ is independent of M more tractable than others, so this restricted class of models is useful for gaining intuition and illustrating different types of behavior.

The more general equation for the evolution of $n(M,t)$ is

$$\begin{aligned}\frac{\partial n(M,t)}{\partial t} &= -\frac{1}{t_s}\frac{\partial(nM\langle\dot{m}(M,t)\rangle)}{\partial M} + C(M,t) - D(M,t)\\ &+ \int_0^M dM' n(M-M',t)n(M',t)\Gamma(M-M',M',t)\\ &- n(M,t)\int_0^\infty dM' n(M',t)\Gamma(M,M',t)\ ,\end{aligned} \tag{12}$$

where the last two terms represent formation of black holes of mass M by mergers of black holes of mass M' and $M - M'$ and loss of black holes of mass M by mergers with other black holes, and the creation (C) and destruction (D) terms allow for processes that are neither smooth accretion nor mergers. Equation (12) is essentially the same equation that Murali et al. (2002) use to model the evolution of the galaxy mass function by accretion and mergers (see their equation A1). The genuine destruction of supermassive black holes seems an unlikely prospect, but they could be removed from the population of galactic nuclei by ejection in three-body encounters (e.g., Valtonen et al. 1994), which would effectively count as destruction for our purposes. We will not consider this possibility in our models here, but it could be a significant effect if there is a long delay between the merger of galaxies and the merger of their central black holes (see, e.g., Madau et al. 2003). We will also assume that there is no black hole creation in the mass and redshift range that we consider, i.e., all of the black holes are already present at the highest redshift in our calculations, and their masses change only by accretion and mergers. A calculation that included the formation of "seed" black holes would need to incorporate the creation term $C(M,t)$, but here we treat the mass distribution of these seeds as the initial condition for evolution of $n(M)$. In the context of our models, a "seed" black hole is one that forms at low efficiency by a process that is not well described by the same $p(\dot{m})$ governing most accretion driven growth — e.g., by the collapse of a supermassive star, a relativistic gas disk, or a relativistic star cluster.

If the accretion, creation, and destruction terms are ignored, equation (12) becomes the "coagulation equation," which has been widely studied in the context of planetesimal growth (see Lee 2000 and references therein) and applied on occasion to star formation and galaxy clustering (e.g., Silk & White 1978; Silk & Takahashi 1979; Murray & Lin 1996; Sheth 1998). Unfortunately, the coagulation equation has analytic solutions only for rather specialized forms of the collision rates $\Gamma(M,M',t)$ that do not seem particularly applicable to black hole evolution, and even these solutions no longer apply once accretion is also important. The usefulness of equation (12) is therefore largely conceptual, and as a basis for numerical calculations given some physically motivated forms of the collision rates. In this paper, we will restrict our consideration of mergers to idealized recipes that are, we hope, sufficient to illustrate their generic effects.

3. Luminosity Functions

3.1. Power-Law $\mathbf{p}(\dot{m})$ for $\dot{m}$ in the Thin-Disk Range

The functions $p(\dot{m}|M)$ and $n(M)$ could in principle have complicated forms, but it proves useful to consider some simple forms and determine qualitative behaviors that would hold true in more general cases. We begin by implementing our framework in a specific case where we consider only the range of accretion rates that corresponds to thin-disk accretion, i.e. $\dot{m}_{\rm crit} < \dot{m} < 1$ and thus $\epsilon_{0.1} = 1$. We also assume that the probability that a black hole accretes at a rate $\dot{m}$ is independent of its mass, i.e. $p(\dot{m}|M) = p(\dot{m})$. We first consider a truncated power-law,

$$p(\dot{m}) = p_* \dot{m}^a, \quad \dot{m}_{\rm min} < \dot{m} < \dot{m}_{\rm max}. \tag{13}$$

[There is also a $\delta-$function at $\dot{m} = 0$, representing inactive black holes, so that $p(\dot{m})$ integrates to one.] In this section, we adopt $\dot{m}_{\rm min} = \dot{m}_{\rm crit} = 0.01$ (see §2.1) and $\dot{m}_{\rm max} = 1$. We adopt a broken power-law form for the black hole mass function,

$$n(M) = \begin{cases} n_* \left(\frac{M}{M_*}\right)^\alpha & M < M_*, \\ n_* \left(\frac{M}{M_*}\right)^\beta & M > M_*. \end{cases} \tag{14}$$

Henceforth, we will frequently refer to the normalization of $n(M)$ in terms of $n_* M_*$, the number density of objects within a logarithmic interval around M_*.

The convolution integral (8) for the luminosity function breaks into three regimes because the range of accretion rates is finite. Luminosities $L < \epsilon_{0.1}\dot{m}_{\rm min} l M_* = 0.01 l M_*$ cannot be generated by black holes with masses $M > M_*$ because they would require an accretion rate below $\dot{m}_{\rm min}$. Thus, only the low mass end of the black contributes to this luminosity regime,

$$\Phi(L) = \int_{\dot{m}_{\rm min}}^{\dot{m}_{\rm max}} p_* \dot{m}^a n_* \left(\frac{L}{\epsilon_{0.1}\dot{m} l M_*}\right)^{\alpha} \frac{1}{\epsilon_{0.1} l \dot{m}} d\dot{m}, \qquad L < 0.01 l M_*. \tag{15}$$

Similarly, luminosities $L > \epsilon_{0.1}\dot{m}_{\rm max} l M_* = l M_*$ cannot be generated by black holes with masses $M < M_*$, so the integral for this regime is the same as equation (15) but with α replaced by β. For the intermediate regime, black holes with masses above and below M_* contribute. It is convenient to define the accretion rate $\dot{m}_{\rm L}$ at which an M_* black hole has luminosity L,

$$\dot{m}_{\rm L} \equiv \frac{L}{\epsilon_{0.1} l M_*}\ , \tag{16}$$

so that the $\Phi(L)$ integral breaks into pieces contributed by the high and low ends of the black hole mass function:

$$\begin{aligned} \Phi(L) &= \int_{\dot{m}_{\rm min}}^{\dot{m}_{\rm L}} p_* \dot{m}^a n_* \left(\frac{L}{\epsilon_{0.1}\dot{m} l M_*}\right)^{\beta} \frac{1}{\epsilon_{0.1} l \dot{m}} d\dot{m} \\ &+ \int_{\dot{m}_{\rm L}}^{\dot{m}_{\rm max}} p_* \dot{m}^a n_* \left(\frac{L}{\epsilon_{0.1} l \dot{m} M_*}\right)^{\alpha} \frac{1}{\epsilon_{0.1} l \dot{m}} d\dot{m}, \qquad 0.01 l M_* < L < l M_*. \end{aligned} \tag{17}$$

The solution for the QLF is

$$\Phi(L) = \begin{cases} \frac{n_* p_*}{l} \frac{1-0.01^{a-\alpha}}{(a-\alpha)} \left(\frac{L}{l M_*}\right)^{\alpha} & L < 0.01 l M_*, \\ \frac{n_* p_*}{l} \left[\frac{\beta-\alpha}{(a-\beta)(a-\alpha)} \left(\frac{L}{l M_*}\right)^{a} - \frac{0.01^{a-\beta}}{a-\beta} \left(\frac{L}{l M_*}\right)^{\beta} + \frac{1}{a-\alpha} \left(\frac{L}{l M_*}\right)^{\alpha}\right] & 0.01 l M_* < L < l M_*, \\ \frac{n_* p_*}{l} \frac{1-0.01^{a-\beta}}{(a-\beta)} \left(\frac{L}{l M_*}\right)^{\beta} & L > l M_*, \end{cases} \tag{18}$$

where the values $\dot{m}_{\rm min} = 0.01$, $\epsilon_{0.1} = 1$, and $\dot{m}_{\rm max} = 1$ have been explicitly included. Note that we have scaled luminosities to the Eddington luminosity $l M_*$ of an M_* black hole.

The most comprehensive observational analysis of the optical quasar luminosity function at $z \lesssim 3$ is that of Boyle et al. (2000), based on the 2dF Quasar Redshift Survey and the Large Bright Quasar Survey. They find that the rest-frame B-band luminosity function can be adequately fit by a double power-law function of the form

$$\Phi(L_B) \propto \left[\left(\frac{L_B}{L_{\rm brk}}\right)^{\alpha_L} + \left(\frac{L_B}{L_{\rm brk}}\right)^{\beta_L}\right]^{-1}, \tag{19}$$

with the break luminosity $L_{\rm brk}$ evolving with redshift. (We denote the break luminosity $L_{\rm brk}$ rather than L_* to keep clear the distinction between this observational quantity and the characteristic parameters of our models, M_* and $\dot{m}_*$.)

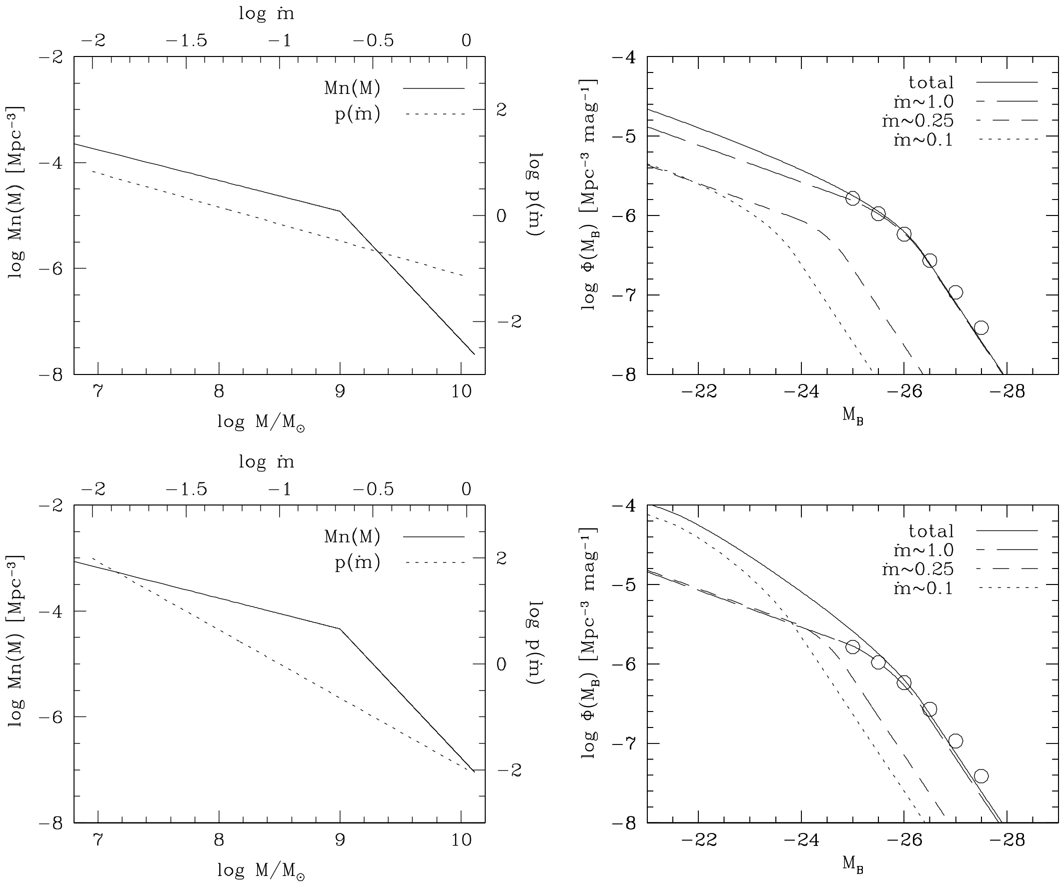

Fig. 1.— QLF for models with a double power-law black hole mass function and a single power-law $p(\dot{m})$, with thin-disk accretion only. Left panels show the input $p(\dot{m})$ (dotted line, top and right axis labels, dimensionless) and mass function (solid line, bottom and left axis labels). In all Figures, we plot $Mn(M)$, the number per comoving Mpc^3 per $\ln M$ interval, rather than the mass function $n(M)$ itself. Right panels show the corresponding B-band luminosity function, plotted in number per comoving Mpc^3 per magnitude against B-band absolute magnitude. Solid lines show the total luminosity function, while long-dashed, short-dashed, and dotted lines show the contributions from accretion rates in the ranges $0.01 < \dot{m} < 0.1$, $0.1 < \dot{m} < 0.25$, and $0.25 < \dot{m} < 1.0$, respectively. Open circles show the double power-law fit to the Boyle et al. (2000) QLF measurements at $z = 2$, over the range of absolute magnitudes probed by the data. Top and bottom rows show results for $p(\dot{m})$ power-law slopes $a = 0$ and $a = -1.2$, respectively, for the same black hole mass function. This and all subsequent figures assume an $\Omega_m = 1$ cosmology with $h = 0.5$.

Figure 1 shows QLFs computed by equation (18) for two different choices of the $p(\dot{m})$ index, $a = -1$ (top) and $a = -2$ (bottom). Left hand panels show the input mass functions and $p(\dot{m})$, and right hand panels show the corresponding QLF. Instead of the mass function $n(M)$ itself, we plot $Mn(M)$, the space density of black holes per $\ln M$ interval, since this quantity is easier to interpret. We adopt $n(M)$ slopes $\alpha = -1.5$ and $\beta = -3.4$ to match the low and high end slopes of the Boyle et al. (2000) QLF. Note that the slopes of the QLFs in Figure 1 are actually $\alpha+1$ and $\beta+1$ because they are plotted in terms of magnitudes rather than luminosities. The QLFs have been converted from bolometric emission into absolute B-band magnitudes by using $L_{\rm bol}/\nu_B L_{\rm B} = 10.4$ (Elvis et al. 1994) to calculate the B-band luminosity and then using $M_B = -2.5(\log\, L_B - 32.67) + 5.48$ to convert the luminosity into an absolute magnitude, where 5.48 is the absolute magnitude of the sun in the B band and $10^{32.67}$ erg s^{-1} is the solar luminosity in the B band. However, it is important to keep in mind that these are scaled bolometric luminosity functions and correspond to true optical luminosity functions only if this $L_{\rm bol}/\nu_B L_{\rm B}$ ratio is constant from quasar to quasar (see §3.5). The points in the right hand panels correspond to Boyle et al.'s fit (eq. 19) at $z \sim 2$, and they cover only the range of magnitudes actually observed at this redshift.

The amplitude of the QLF is directly proportional to p_* and to the mass function normalization n_*. The value of p_* determines the probability that a given black hole will be "on" at some non-zero luminosity. For the $a = -2$ model (lower panel), we have somewhat arbitrarily chosen p_* so that this probability, $\int_{\dot{m}_{\rm min}}^{\dot{m}_{\rm max}} p(\dot{m})d\dot{m}$, is equal to one. Though all of the black holes are active in this case, most of the activity is at accretion rates near $\dot{m}_{\rm min}$. We then choose a mass function normalization $n_* M_* = 4.6 \times 10^{-5}{\rm Mpc}^{-3}$ so that the model QLF matches the Boyle et al. (2000) data point at the break luminosity. For the $a = -1$ model, we have kept the same $n(M)$ and again chosen p_* to match the observed QLF at the break. Note that equation (18) reveals a complete degeneracy between n_* and p_* in the QLF. Thus, at a fixed time, it is impossible to determine from the QLF alone if there is a high number density of black holes of which a small fraction are accreting or a low number density of black holes of which a large fraction are accreting. We will see later that the evolution of the QLF breaks this degeneracy to some degree.

The QLF in the upper right panel of Figure 1 demonstrates one of the important general features of this solution: the slopes of the low and high luminosity ends of the QLF are determined by the slopes of the low and high mass ends of the black hole mass function (see eq. 18). This behavior follows whenever $p(\dot{m})$ is independent of M and the range of accretion rates is finite. This result can be understood by considering that a small range $d\dot{m}$ at $\dot{m} = 1$ gives $\Phi(L) = n(M)(\epsilon_{0.1}l)^{-1}p(\dot{m} = 1)d\dot{m}$ with $M = L/(\epsilon_{0.1}l)$, which is a simple mapping of the black hole mass function. The same range $d\dot{m}$ at a lower $\dot{m}$ gives a contribution to the QLF mapped to luminosities fainter by a factor of $\dot{m}$ and up or down by a factor proportional to $p(\dot{m})/p(\dot{m} = 1)$. The total QLF is just the sum of these transformed black hole mass functions. Since we have low and high luminosity regimes in which only one slope of $n(M)$ contributes, the sum in these regimes will be a sum of power-laws with the same slope, resulting in a QLF with the same slope. The mid-range luminosity is more complicated because both parts of the black hole mass function are contributing.

The relative contributions from the low and high end of the black hole mass function depend on the relative probability of low and high accretion rates. This can be seen in the middle part of equation (18), which shows that the slope of the probability distribution as well as the slopes of $n(M)$ affect the shape of the QLF in this luminosity regime. The agreement of asymptotic slopes between $n(M)$ and $\Phi(L)$ motivates our choice of a double power-law $n(M)$, though we will see in §5.2 that this choice has some problems when compared to local observations.

The $a = -1$ model in Figure 1 corresponds to equal probabilities of accretion in each logarithmic interval of $\dot{m}$ between $\dot{m}_{\rm min}$ and $\dot{m}_{\rm max}$, but the emissivity of the population is dominated by accretion rates close to $\dot{m}_{\rm max}$ because luminosities are proportional to $\dot{m}$. The model QLF is in reasonably good agreement with the Boyle et al. (2000) data. Dotted, short-dashed, and long-dashed lines in the QLF panels show the contribution from accretion rates in the ranges $0.01 < \dot{m} < 0.1$, $0.1 < \dot{m} < 0.25$, and $0.25 < \dot{m} < 1.0$ respectively. For $a = -1$, the QLF is dominated by black holes with near-Eddington accretion rates at all luminosities. The curves for lower accretion rates are shifted horizontally to lower luminosities, with slight vertical shifts because the three bins are not equal logarithmic intervals.

The $a = -2$ model has equal contributions to the emissivity from each logarithmic interval of $\dot{m}$. The slow transition between the low and high luminosity regimes yields a worse fit to the Boyle et al. (2000) QLF. High accretion rates still dominate the high end of the QLF, but low accretion rates dominate the low end. In general, low $\dot{m}$ black holes can more easily make a significant contribution below the break in the luminosity function because a shift "left" can be more easily compensated by a shift "up," especially when the slope of the low end of the mass function is shallow. However, above the break in luminosity, it is difficult for low $\dot{m}$ accretors to make a contribution. For example, a $p(\dot{m})$ slope $a \sim -3.3$ would be required to make the contribution of low and high accretion rates comparable at high luminosities in Figure 1.

3.2. Double Power-Law p($\dot{m}$) with ADAF and Super-Eddington Accretion

Our assumptions for this section are similar to those of §3.1, except that we consider a wider range of allowed accretion rates and adopt a double power-law form of $p(\dot{m})$. Specifically, we consider accretion rates in the range $\dot{m}_{\rm min} = 10^{-4} < \dot{m} < \dot{m}_{\rm max} = 10$, which allows for ADAF, thin-disk, and super-Eddington accretion modes, for which we adopt the efficiencies given in equation (7). The functional form of $n(M)$ we consider here is the double power-law expressed in equation (14), and the form of $p(\dot{m})$ is analogous,

$$p(\dot{m}) = \begin{cases} p_* \left(\frac{\dot{m}}{\dot{m}_*}\right)^a & \dot{m} < \dot{m}_*, \\ p_* \left(\frac{\dot{m}}{\dot{m}_*}\right)^b & \dot{m} > \dot{m}_*, \end{cases} \tag{20}$$

where $\dot{m}_*$ is the characteristic accretion rate at the break in $p(\dot{m})$. The solution for the QLF in this case is given in Appendix A.

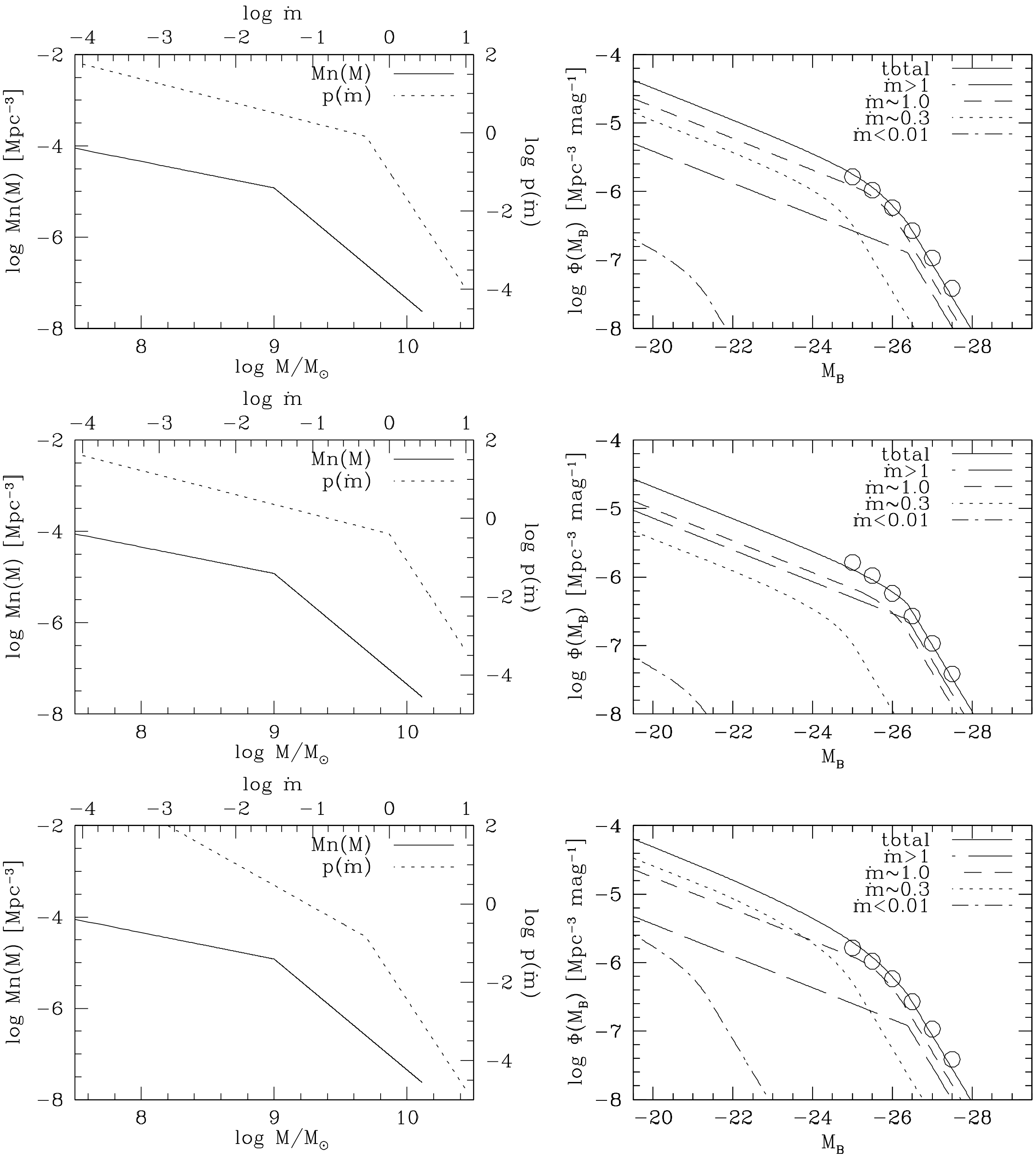

Fig. 2.— Like Fig. 1, but for double power-law $p(\dot{m})$ covering a wider range of accretion rates, including values in the ADAF regime and super-Eddington regime. Efficiencies as a function of $\dot{m}$ are calculated according to eq. (7). In the right panels, solid lines show the total QLF, and other lines show the contributions from the ranges $\dot{m} > 1$ (super-Eddington, long-dashed), $0.3 < \dot{m} < 1$ (thin-disk, high accretion rate, short-dashed), $0.01 < \dot{m} < 0.3$ (thin-disk, low accretion rate, dotted), and $10^{-4} < \dot{m} < 0.01$ (ADAF, dot-dashed). Relative to the case in the top row, the middle and bottom rows show cases with a higher characteristic accretion rate ($\dot{m}_* = 1$ vs. $\dot{m}_* = 0.5$) and a steeper slope at low accretion rates ($a = -1.1$ vs. $a = -0.5$), respectively.

Figure 2 is analogous to Figure 1, but for the double power-law $p(\dot{m})$. In the upper panels, we adopt $a = -0.5$, $b = -3$, and $\dot{m}_* = 0.5$. At high luminosity, the QLF is dominated by the higher accretion rates in the thin-disk mode, but there is also a significant contribution from the super-Eddington mode. At low luminosity, the contribution from higher accretion rates in the thin-disk mode dominates over lower accretion rates in the thin-disk mode, with the super-Eddington mode becoming much less significant. The contribution of the ADAF mode to the QLF is barely noticeable, and this is typically the case in our models because low $\dot{m}$ and low $\epsilon_{0.1}$ combine to push black hole luminosities down to $\sim 10^{-2} - 10^{-6}$ of Eddington, which is usually below observed luminosities. ADAF contributions can be more significant at low redshifts and X-ray wavelengths, as discussed in §3.5 below.

The middle panel shows the effect of increasing $\dot{m}_*$ to 1.0, with p_* decreased to keep $\Phi(L_{\rm brk})$ fixed. Super-Eddington accretion is now common enough that it dominates the high end of the QLF, with a comparable but smaller contribution from thin-disk accretion with $0.3 < \dot{m} < 1.0$. Lower luminosities are dominated by thin-disk accretion, with high accretion onto lower mass black holes more important than low accretion onto high mass black holes.

The bottom panels show a case with $\dot{m}_* = 0.5$ and a low end slope of $a = -1.1$, which weights $p(\dot{m})$ more strongly to low accretion rates. The high luminosity end of the QLF in the bottom panel has contributions similar to those in the upper panel, but the low luminosity end is now dominated by lower accretion rates in the thin-disk mode. The turnover of the QLF in the lower panel is less sharp than in the upper panel because of the change of the relative fraction of high and low accretion rates contributing to the QLF over the transitional range of luminosity. The contribution from the ADAF mode is higher than in the top panel, but it remains small at all luminosities.

As noted earlier, with the luminosity function alone there is a complete degeneracy between the normalizations n_* and p_*, subject only to the limitation that the fraction of black holes accreting at a given time not exceed 100%. There is also a partial degeneracy between the characteristic values M_* and $\dot{m}_*$: increasing M_* shifts the break in the QLF to higher luminosity, but reducing $\dot{m}_*$ makes lower accretion rates dominate the QLF, shifting the break back down. With our assumption that the efficiency $\epsilon_{0.1}(\dot{m})$ decreases for $\dot{m} > 1$, this tradeoff is limited to a factor of a few, since changes in $\dot{m}_*$ also affect the shape of the turnover in the QLF. We will see later that this tradeoff is even more restricted when the evolution of the QLF is considered.

3.3. Mass Distribution of Active Black Holes

The masses of active black holes can be estimated using reverberation mapping (e.g., Wandel et al. 1999; Onken & Peterson 2002), the widths of lines such as Hβ and CIV (e.g., Laor 1998; Vestergaard 2004; Corbett et al. 2003), or indirectly from the properties of host galaxies (e.g., Dunlop et al. 2003), the variability power spectrum (Czerny et al. 2001), or the spectral energy distribution (Kuraszkiewicz et al. 2000). The distribution of active black hole masses at a given luminosity depends on both the underlying black hole mass function $n(M)$ and the distribution of accretion rates $p(\dot{m})$. While the necessary measurements are challenging, especially at high redshift, they can play a critical role in discriminating among models that make very similar predictions for the luminosity function.

If there is a maximum value of the product $\epsilon_{0.1}\dot{m}$ (e.g., unity if luminosities cannot exceed Eddington, as in eq. 7), then black holes with $M < (L/l)[(\epsilon_{0.1}\dot{m})_{\rm max}]^{-1}$ cannot contribute to the QLF at luminosity L because they cannot shine brightly enough. Conversely, black holes with $M > (L/l)[(\epsilon_{0.1}\dot{m})_{\rm min}]^{-1}$ are always brighter than L, when they are active at all. The contribution to the QLF from black holes with masses in the allowed range is the product of the number density of black holes with mass M and the probability of having an accretion rate in the range $\dot{m} \to \dot{m}+\Delta\dot{m}$ that yields luminosity $L \to L+\Delta L$. Thus, for a quasar of luminosity L, the relative probability that its black hole has mass M_1 or M_2, if both masses are in the allowed range, is

$$\frac{p(M_1|L)}{p(M_2|L)} = \frac{n(M_1)p\left(\dot{m}=\frac{L}{\epsilon_{0.1}lM_1}\right)\frac{\Delta L/M_1}{\epsilon_{0.1}l}}{n(M_2)p\left(\dot{m}=\frac{L}{\epsilon_{0.1}lM_2}\right)\frac{\Delta L/M_2}{\epsilon_{0.1}l}} = \left(\frac{M_1}{M_2}\right)^{\alpha-(a+1)}, \tag{21}$$

where $(\Delta L/\epsilon_{0.1}lM) = \Delta\dot{m}$. The rightmost equality applies for constant $\epsilon_{0.1}$ and power-law forms of $p(\dot{m})$ and $n(M)$, with slopes of a and α respectively. For this case we see that if $\alpha < a+1$ then the rising low end of the mass function results in low mass black holes with high accretion rates dominating the active population at luminosity L. Conversely, if $\alpha > a+1$, then low accretion rates are common enough that high mass black holes with low $\dot{m}$ dominate. A single power-law can only approximate $n(M)$ or $p(\dot{m})$ over a finite range, but this example gives insight into the more general case and allows one to judge whether quasars of luminosity L are likely to be dominated by masses near a break in $n(M)$, or by accretion rates near a break in $p(\dot{m})$ or $\epsilon_{0.1}(\dot{m})$. For example, the high luminosity regime Figure 2 is dominated by Eddington luminosity black holes (near a break in $\epsilon_{0.1}$) because of the steep slope of $n(M)$ at high masses.

For a statistical quantity that is easier to measure, it is usually desirable to integrate equation (21) to obtain the distribution of black hole masses for a specified range in luminosities. We will show in §5 that such statistics can discriminate between models that yield similar QLFs over the range of observed luminosities but have significantly different parameter values.

3.4. Mass Dependence of p($\dot{m}$)

A general mass-dependent distribution of accretion rates can be written in the form $p(\dot{m}|M) = p_0(\dot{m})D(M|\dot{m})$, where $p_0(\dot{m}) = p(\dot{m}|M_0)$ at an arbitrarily chosen mass scale M_0 and the function D encodes the mass dependence, with $D(M_0|\dot{m}) \equiv 1$. To understand the potential influence of mass dependence, we will consider the restricted case in which the function $D(M)$ is independent of $\dot{m}$, and $p(\dot{m}|M)$ is thus a separable function. In this class of models, the relative probabilities of high and low accretion rates are independent of mass, but the overall duty cycle can have an arbitrary mass dependence. Recall that the general expressions for the QLF (eq. 8) and the distribution $p(M|L)$ of black hole masses at a given luminosity (eq. 21) involve $p(\dot{m}|M)$ and $n(M)$ only through the product $p(\dot{m}|M)n(M)$. Therefore, for any QLF and $p(M|L)$ generated by a black hole mass function $n(M)$ and a mass-independent $p(\dot{m})$, there is a family of models with mass function $n'(M) = n(M)/D(M)$ and mass-dependent accretion rate distribution $p'(\dot{m}|M) = p(\dot{m})D(M)$ that predicts the same QLF and $p(M|L)$, for any choice of the function $D(M)$. The mass dependence of $p(\dot{m})$ therefore introduces a rather serious degeneracy into models of the luminosity function, which cannot be broken by measurements of the distribution of *active* black hole masses.

The key difference within this class of degenerate models is the relation between the luminosity function and the underlying black hole mass function $n(M)$. In particular, we have shown in §3.1 and §3.2 that for a mass-independent $p(\dot{m})$ the low and high luminosity slopes of the QLF match the low and high mass slopes of $n(M)$. This is no longer the case if $p(\dot{m})$ depends on M. For example, with $D(M) \propto M^x$ and a double power-law $n(M)$, the QLF has asymptotic slopes $\alpha + x$ and $\beta + x$, rather than α and β. Thus, a measurement of the mass function of all black holes, not just the active ones, is crucial to diagnosing the mass dependence of accretion rates.

Our discussion above focuses on the QLF at a particular redshift, and the degeneracy applies if $p(\dot{m}|M)$ and $n(M)$ can be chosen at will. If one considers a range of redshifts over which black holes grow by a substantial factor, then the predictions of models with different mass dependence of $p(\dot{m}|M)$ are likely to diverge, since the mass-dependence of growth rates will change the shape of $n(M, z)$ from one model to the next (see §4.2.2). Thus, this degeneracy should be less serious in a complete evolutionary model of the population. Furthermore, a measurement of $n(M)$ at $z = 0$ may be sufficient to diagnose the mass dependence of $p(\dot{m}|M)$ at higher redshift.

3.5. Wavelength Dependent Efficiencies

Equation (8) gives the bolometric luminosity function $\Phi(L)$ in terms of the accretion rate distribution, the black hole mass function, and the efficiency $\epsilon_{0.1}$, which may itself be a function of $\dot{m}$. If all quasars have the same spectral energy distribution (SED), then the translation to the luminosity function in a band at frequency ν is straightforward:

$$\Phi(L_\nu) = \int_{\dot{m}_{\rm min}}^{\dot{m}_{\rm max}} p(\dot{m}) n \left(M = \frac{L_\nu}{\epsilon_{0.1}\dot{m} l F_\nu} \right) \frac{1}{\epsilon_{0.1}\dot{m} l F_\nu} d\dot{m} \ , \qquad (22)$$

where $F_\nu \equiv L_\nu / L_{\rm bol}$ is the fraction of the quasar's bolometric luminosity that emerges in the ν-band. (Note that we are using subscript-ν to represent a finite band, not a monochromatic flux density.) We have so far presented results for the rest-frame B-band luminosity function, assuming that all accreting black holes have the broad-band SED estimated by Elvis et al. (1994). For a universal SED, the luminosity functions in all bands are just shifted versions of the bolometric luminosity function, so they all have the same shape.

The story is more interesting if some accreting black holes have radically different SED shapes.

Here we will consider two representative examples, an "obscured" accretion mode in which optical, UV, and soft X-ray radiation are absorbed by gas and dust near the nucleus and re-radiated in the far-IR, and an ADAF mode that has a high ratio of X-ray flux to optical flux (in addition to a reduced bolometric efficiency). Obscured accretion is thought to play an important role in producing the X-ray background, and typical synthesis models in the literature have a $\sim 4:1$ ratio of obscured to unobscured sources (e.g., Comastri et al. 1995; Fabian & Iwasawa 1999). More recent results from *Chandra* show that the obscured fraction is probably lower than this, especially at high luminosities (e.g., Barger et al. 2002; Ueda et al. 2003). ADAFs are expected on theoretical grounds to have depressed UV/optical emission relative to X-ray and far-IR (Narayan et al. 1998), and many low luminosity AGN in the nearby universe, including Sgr A* in the Galactic Center, appear to have these broader SED shapes (e.g., Ho 1999). There are, of course, other possibilities for SED variations, including a steady change of SED shape with black hole mass caused by the lower characteristic temperatures around higher mass black holes, a change of SED shape in the super-Eddington regime, and perhaps a transition within the $\sim 0.1-1L_{\rm Edd}$ regime as the accretion disk grows in importance relative to the hard X-ray corona.

The first task is to calculate values of F_ν for the model SEDs. We will consider luminosity functions in the rest-frame B-band, soft (0.5-2 keV) and hard (2-10 keV) X-ray bands, and a "far-IR" band that we take to cover the range $10-1000\mu$m. Since most X-ray studies work with observed-frame fluxes (though see Cowie et al. [2003], Steffen et al. [2003], and Ueda et al. [2003] for recent efforts to measure evolution of the rest-frame 2-8 keV luminosity function), we also consider the soft and hard X-ray bandpasses redshifted to $z = 0.5$, 1, and 2. We assume that our far-IR band is wide enough that all high-energy radiation absorbed in obscured systems is re-radiated somewhere within it. Unfortunately, realistic experiments are likely to probe a narrower band, for which the predictions may be quite sensitive to assumptions about dust temperatures and departures from a blackbody spectrum. We will not consider radio luminosities here. To the extent that the radio-quiet/radio-loud dichotomy is a simple effect of orientation or black hole spin, it could be treated as a stochastic variation analogous to our treatment of obscuration. However, radio luminosity may also be connected to accretion rate or to black hole mass (Dunlop et al. 2003).

Table 1 summarizes our values of F_ν. We assume that unobscured quasars with accretion rates $\dot{m} > \dot{m}_{\rm crit} = 0.01$ have the mean radio quiet SED of Elvis et al. (1994), and we obtain F_ν by integrating over the appropriate wavelength bands. Simple power-law extrapolations were used in regions without observations, and the high energy SED was extended to 30 keV by using a power-law with photon index $\Gamma_X = 1.9$. For obscured quasars, we assume the same "underlying" SED and compute absorption in the X-rays by taking a mean X-ray photon index of $\Gamma_X = 1.89$ and an obscuring column density of $N_H = 3 \times 10^{23}$ cm^{-2}, which represents the median value used in the X-ray background sythesis models of Comastri et al. (2001). We further assume that the high gas column is accompanied by enough dust to completely extinguish the B-band flux; this assumption may be inaccurate, as some observed systems appear to have significant optical/UV flux despite strong absorption in the X-ray. Finally, we assume that all of the energy absorbed from the optical to the soft X-rays, about 52% of the bolometric energy in the Elvis et al. (1994) SED, is reradiated in the FIR. This assumption seems physically plausible, though it is also possible that some or most of the absorbed energy is channeled into driving an outflow and never radiated at all.

As in our previous calculations, we assume that accretion rates $\dot{m} < \dot{m}_{\rm crit} = 0.01$ lead to reduced efficiency, ADAF flows, but now we assign these flows a different SED. The appropriate SED for ADAF systems is quite uncertain, and we have elected simply to take the nucleus of M81 as representative of all black holes accreting in this mode. We calculate F_ν values from the observations tabulated in Ho (1999, tables 2 and 9 and §2.1), using the estimated $L_{\rm bol}$ and directly observed L_ν when available and otherwise using the observed monochromatic νL_ν values and an appropriate α_ν to integrate over the waveband of interest. Since the observed FIR flux value of M81 is an upper limit, we used the model of M81 presented in Quataert et al. (1999, fig. 1) to estimate $F_{\rm FIR}$ for the ADAF mode.

With the results of Table 1 in hand, we can again calculate multi-wavelength luminosity functions using equation (22), but now the results must be computed separately for the three modes (thin-disk, obscured, ADAF) and added together. We assume that super-Eddington accretion has a thin-disk SED; this seems unlikely, but we do not have much idea of what to assume in its stead, and it makes little difference to our results.

Figure 3 illustrates the potential influence of a large obscured quasar population on the multi-wavelength QLF at $z = 2$. Solid lines show results for a model with no obscured quasars, with the same black hole mass function and $p(\dot{m})$ used in upper panels of Figure 2. Dashed lines show a model in which these unobscured quasars are only 20% of the total — i.e., we multiply $p(\dot{m})$ by four but assign 80% of the systems with $\dot{m} > 0.01$ the obscured SED of Table 1. Since the obscured SED has no B-band flux, the B-band luminosity function is unchanged (top left panel). However, obscuration in the observed-frame 2-10 keV band (rest-frame 6-30 keV) is minimal, so the hard X-ray luminosity function is nearly a factor of five higher in amplitude, with the contribution

Table 1. Fractional Bolometric Output

	F_{2-10}	F_{3-15}	F_{4-20}	F_{6-30}	$F_{0.5-2}$	$F_{0.75-3}$	F_{1-4}	$F_{1.5-6}$	$F_{B-\rm band}$	$F_{\rm FIR}$
Thin-Disk	0.026	0.028	0.030	0.031	0.020	0.020	0.020	0.022	0.025	0.17
Obscured	0.008	0.015	0.020	0.027	1×10^{-9}	1.3×10^{-5}	2.9×10^{-4}	2.6×10^{-3}	0.00	0.79
ADAF	0.084	0.089	0.093	0.099	0.057	0.061	0.064	0.068	0.012	0.03

Note. — Values of the inverse bolometric correction, $F_{\nu-\rm band}$, where $L_{\nu-\rm band} = F_{\nu-\rm band} L_{\rm Bol}$. Values include intrinsic X-ray ranges at $z = 2, 1$, and 0.5 that correspond to the observed soft and hard bands at $z = 0$, as well as the rest-frame B-band and FIR band.

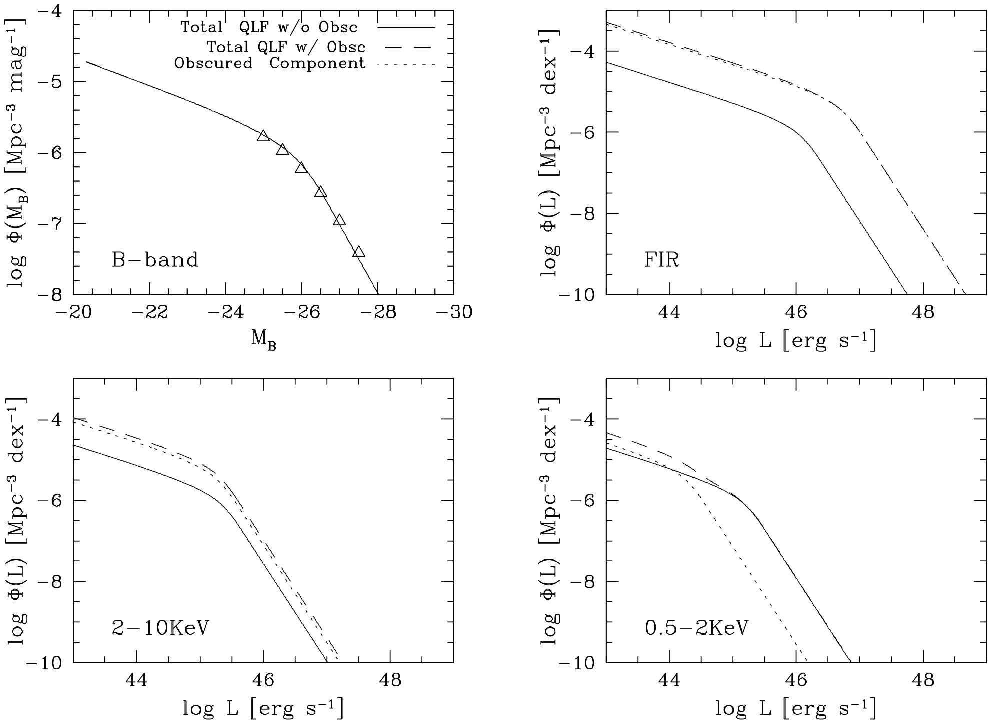

Fig. 3.— Potential influence of obscuration on multi-wavelength QLFs at $z \sim 2$. Solid lines show B-band, FIR, 0.5-2 keV, and 2-10 keV luminosity functions of a model in which all quasars have the Elvis et al. (1994) SED, with $p(\dot{m})$ and $n(M)$ chosen to match the Boyle et al. (2000) B-band results (triangles in upper left). Dashed lines show QLFs for a model in which 20% of quasars have the Elvis et al. (1994) SED and 80% have an obscured SED corresponding to $N_H = 3 \times 10^{23}\mathrm{cm}^{-2}$ (see F_ν values in Table 1). Dotted lines show the contribution of obscured systems in this model (note that we assume complete optical obscuration, hence no contribution in B). X-ray luminosities are *observed-frame* at $z = 2$.

of obscured quasars (dotted line) dominating at all luminosities. For the 0.5-2 keV band (rest-frame 1.5-6 keV) the situation is more complicated, since obscuration suppresses flux in this band by nearly a factor of ten. At high luminosities, unobscured quasars dominate the QLF because the greater numbers of obscured quasars are not enough to compensate for their reduced fluxes. However, the obscured population boosts the faint end of the soft X-ray QLF by about a factor of two, with obscured and unobscured systems making roughly equal contributions.

The most dramatic effect of the obscured population is to boost the FIR luminosity function by a large factor, since with our assumptions the FIR band contains 79% of the bolometric flux of obscured systems but only 17% of the bolometric flux of unobscured systems. The combination of more systems and more flux per system boosts $\Phi(L_{\rm FIR})$ by almost two orders of magnitude at high luminosities, with the FIR QLF totally dominated by obscured systems at every redshift. Note that our treatment of $F_{\rm FIR}$ implicitly assumes that obscured and unobscured systems are two distinct populations, one with high gas columns and one without. It is also possible, as assumed by Sazonov, Ostriker, & Sunyaev (2003), that the difference between obscured and unobscured systems is simply one of orientation, and that even systems that are unobscured along our line of sight have most of their optical, UV, and soft X-ray emission absorbed by a dusty torus and re-radiated isotropically in the FIR. In this case, $F_{\rm FIR}$ would be essentially the same for both populations, so at a given FIR luminosity obscured and unobscured systems would be represented in their global ratio (i.e., 4:1 in our model). The joint FIR-optical or FIR-soft X-ray luminosity functions would distinguish these two scenarios.

Figure 3 shows that a large population of obscured quasars can substantially alter the relation between B-band, X-ray, and FIR luminosity functions, as one would expect. The difference in FIR would persist at all redshifts. At low redshifts, on the other hand, the 0.5-2 keV band is almost completely suppressed by a column density $N_H = 3 \times 10^{23} {\rm cm}^{-2}$, and the 2-10 keV band is significantly suppressed, so the effect of an obscured population on the luminosity function in these bands is reduced.

Although we include the ADAF mode in our calculations for Figure 3, we find that ADAF systems make no significant contribution to the QLF at any luminosity likely to be observed at $z \sim 2$, for any plausible choice of our model parameters. However, the situation could be different at low redshift, in part because observations reach to lower luminosities, but even more because (as we discuss in §4.2.1 below) matching the observed QLF evolution requires a shift of $p(\dot{m})$ towards lower characteristic accretion rates at low redshifts, thus giving systems with $\dot{m} < \dot{m}_{\rm crit}$ more chance to compete. Figure 4 shows two models with choices of $p(\dot{m})$ and $n(M)$ that give reasonable matches to the Boyle et al. (2000) B-band QLF at $z \sim 0.5$. The upper panel shows the two $p(\dot{m})$ distributions, which are double power-laws with $\dot{m}_* = 0.03$ and 0.012, respectively. The black hole mass functions (not shown) have corresponding M_* values of $1.21 \times 10^9 M_\odot$ and $2.5 \times 10^9 M_\odot$.

Solid lines in the middle and bottom panels show the predicted luminosity functions for the first model in rest-frame B-band and observed-frame $2-10$ keV, respectively. Total luminosity functions

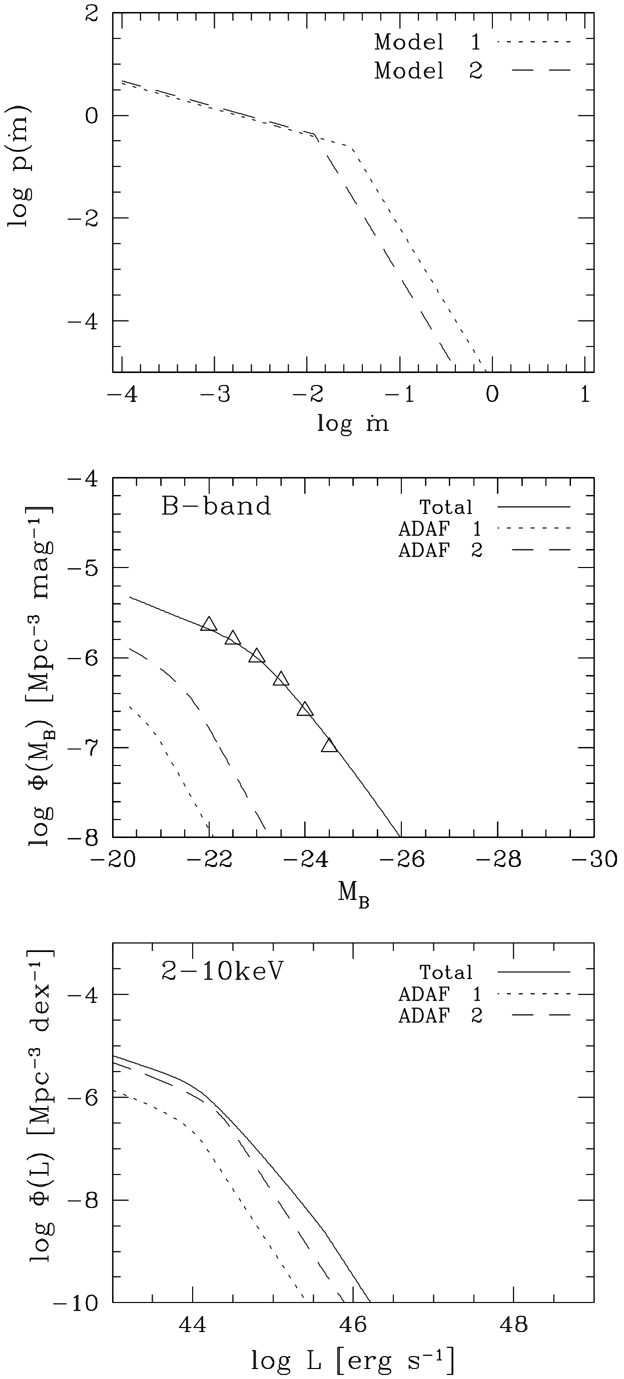

Fig. 4.— Potential influence of ADAFs on multi-wavelength QLFs at $z \sim 0.5$. The top panel shows $p(\dot{m})$ distributions for two models with different $\dot{m}_*$ (the values of M_* are also different). The total QLFs for the two models are nearly identical in both B-band (middle panel) and 2-10 keV (bottom panel); solid lines show the total QLFs of Model 1. Dotted and dashed lines show the contributions of ADAF accretion ($\dot{m} < 0.01$) for the two models, which are small in B-band but substantial at 2-10 keV for Model 2.

for the second model are nearly identical. However, the relative contribution of the ADAF mode, shown by the dotted and dashed lines in these panels, is quite different between the two models and between the two bands. In B-band, ADAF contributions to the luminosity function are strongly suppressed, even for $\dot{m}_* = 0.012$, because of the low optical flux of the ADAF SED. However, the relatively high hard X-ray flux allows ADAF accretion to dominate the low end of the luminosity function for the $\dot{m}_* = 0.012$ model and to make a significant contribution even at high luminosities. In the $\dot{m}_* = 0.03$ model, on the other hand, ADAF accretion is a subdominant contribution to the QLF at all luminosities. Results for the FIR and $0.5 - 2$ keV luminosity functions are qualitatively similar to those for B-band and $2 - 10$ keV, respectively, as one would expect from the values of F_ν in Table 1. Measurements of the optical and X-ray luminosity functions alone would not distinguish the two models shown in Figure 4, but the low-$\dot{m}_*$ model predicts that X-ray selected quasars (largely ADAF systems) should have systematically different SED shapes from optically selected quasars (mostly thin-disk systems), while the high-$\dot{m}_*$ model predicts that thin-disk SEDs dominate both populations.

4. Evolution

We now turn to evolutionary calculations, applying the basic principles of §2.2. We assume that the accretion physics — the dependence of $\epsilon_{0.1}$ and SED shape on accretion rate — is independent of redshift, although the distribution of accretion rates itself evolves. This assumption seems reasonable, since the "microphysics" has no direct knowledge of the age of the universe, but one could imagine that systematic changes in the host galaxy population might affect the influence of gas or dust obscuration on SED shapes, and perhaps even that galaxy mergers could alter the fraction of spinning black holes with higher bolometric efficiencies. With our assumption and a specified model of the accretion physics, $p(\dot{m}|M,z)$ determines the evolution of $\Phi(L)$ in all wavebands, since it both determines the evolution of $n(M)$ and specifies the probability that black holes of a given mass shine at a given luminosity. However, mergers can alter the evolution of the QLF by changing $n(M)$ independently of $p(\dot{m}|M,z)$.

We focus in this section on the optical luminosity function, which is observed to rise by a large factor between $z \sim 5$ and $z \sim 3$ (Warren, Hewett, & Osmer 1994; Schmidt, Schneider, & Gunn 1995; Fan et al. 2001) and decline by a large factor between $z \sim 2$ and $z \sim 0$ (Schmidt 1968; Boyle et al. 2000). At $z \lesssim 2.5$, Boyle et al. (2000) find a break in the luminosity function (eq. 19) that evolves towards lower luminosities at lower redshifts, a form of "luminosity evolution" that cannot be described by a simple vertical shift in amplitude ("density evolution"). We will devote considerable attention to the implications of this result, though we should note that Wolf et al. (2003) reach a more ambiguous conclusion about the need for luminosity evolution, using a data set (COMBO-17) that reaches still lower luminosities. At $z \gtrsim 2.5$, current observations probe mainly the high luminosity end of the QLF, where the data are adequately described by a single power-law.

We begin our discussion below with a few remarks on the definitions of quasar lifetimes and duty cycles. We then present evolutionary calculations for a number of specific models designed to illustrate general points, working within the class of models discussed in §3.2: double power-law $p(\dot{m})$, double power-law $n(M)$, and $\epsilon_{0.1}(\dot{m})$ as defined in equation (7). In §4.2, we consider cases in which accretion alone drives the evolution of $n(M)$, looking first at the declining phase of QLF evolution ($z \leq 2$) and then at the growing phase. In §4.3, we use simple models to illustrate the potential impact of mergers on the evolution of $n(M)$ and the QLF.

4.1. Lifetimes and duty cycles

One of the key elements in models of quasar evolution is the typical quasar lifetime, or, nearly equivalent, the duty cycle of quasar activity (see Martini [2003] for a review of observational estimates). In a simple "on-off" model of the quasar population, where a black hole is either shining at a fixed fraction of its Eddington luminosity or not accreting at all, it is clear what these concepts refer to: at a given redshift, the duty cycle is the fraction of black holes that are active at any one time, and the typical lifetime is the integral of the duty cycle over the age of the universe. However, if $p(\dot{m})$ is broad, and in particular if there is a tail of increasing probability towards low $\dot{m}$, then defining a black hole to be "on" if it is accreting at *any* non-zero rate may not be particularly useful, since changes to $p(\dot{m})$ that have negligible observational effect (such as changing the lower cutoff $\dot{m}_{\rm min}$) may have a large impact on the implied duty cycle or lifetime. Such a definition is also difficult to relate to black hole growth or emissivity.

For our purposes, it is more useful to consider the accretion weighted lifetime

$$t_{\rm acc} \equiv \int_{t_i}^{t_f} \dot{m}(t)\, dt \ . \tag{23}$$

The ratio of $t_{\rm acc}$ to the Salpeter lifetime ($t_s = 4.5 \times 10^7$ yr, cf. eq. 5) gives the number of e-folds of mass growth, i.e.,

$$M_f = M_i \exp\left[\frac{t_{\rm acc}}{t_s}\right] . \tag{24}$$

Typically, the initial time t_i would refer to some time after the formation of "seed" black holes but before the main epoch of mass accretion, and t_f would refer to $z = 0$. However, since we model the declining and growing phases of quasar evolution separately below, we will generally choose t_i and t_f to correspond either to the redshift interval $z = 2 - 0$ or to the redshift interval $z = 5 - 2$. A useful way to characterize the mean accretion rate at a given redshift is

$$t_{{\rm acc},z} \equiv \langle \dot{m} \rangle H^{-1}(z) \ . \tag{25}$$

The ratio $t_{{\rm acc},z}/t_s$ is the number of e-folds of mass growth that would occur if the mean accretion rate were to stay constant for the Hubble time $H^{-1}(z)$.

For constant efficiency $\epsilon_{0.1}$, weighting the lifetime by $\dot{m}$ is equivalent to weighting by $L/(\epsilon_{0.1}L_{\rm Edd})$, so the accretion weighted lifetime is simply related to the luminosity weighted lifetime. Alternatively, one can weight activity by the ratio of a black hole's luminosity $L(t)$ to its *final* Eddington luminosity lM_f. In this case, the weighted lifetime (for constant efficiency and no black hole mergers) is $t_s\epsilon_{0.1}(1-M_i/M_f)$, which approaches the Salpeter lifetime in the limit that the black hole mass grows by a large factor.

4.2. Pure Accretion Driven Evolution

4.2.1. Declining Phase with Mass Independent $p(\dot{m})$

For pure accretion driven evolution, only the first term of equation (12) enters into the evolution of $n(M,t)$. We start by considering the case with $p(\dot{m})$ independent of mass, so that the "self-similar" solution given in equation (11) applies. For a double power-law $p(\dot{m})$, the evolution of the QLF depends on the time evolution of the normalization $p_*(t)$ and the characteristic accretion rate $\dot{m}_*(t)$. We initially consider the declining phase of QLF evolution and take both functions to be power-laws of time:

$$p(\dot{m}|t) = \begin{cases} p_*(t)\left(\frac{\dot{m}}{\dot{m}_*(t)}\right)^a & \dot{m} < \dot{m}_*(t) \\ p_*(t)\left(\frac{\dot{m}}{\dot{m}_*(t)}\right)^b & \dot{m} > \dot{m}_*(t) \end{cases}, \qquad p_*(t) = p_{*,i}\left(\frac{t}{t_i}\right)^{\gamma_p}, \quad \dot{m}_*(t) = \dot{m}_{*,i}\left(\frac{t}{t_i}\right)^{\gamma_m}, \quad (26)$$

where t_i represents the time from which the initial QLF is evolved, and $p_{*,i}$ and $\dot{m}_{*,i}$ correspond to the values of p_* and $\dot{m}_*$ at time $t=t_i$.

For $\gamma_m = 0$, "pure p_* evolution," the relative probability of given accretion rates remains fixed over time, but the overall probability of accreting per unit time declines. This is analogous to "pure density evolution" models of the QLF, though the masses of the black holes continue to evolve. For $\gamma_p = 0$, "pure $\dot{m}_*$ evolution," the relative probabilities of given accretion rates change over time as the break in $p(\dot{m})$ evolves. This is analogous to "pure luminosity evolution," though again, the black hole mass function is evolving along with the evolution of $p(\dot{m})$. Physically, p_* evolution could be connected to a decline in the frequency of galaxy interactions as the universe gets older. Evolution of $\dot{m}_*$ could arise from declining gas fractions of galaxies, or a decline in their ability to funnel gas to the center as they become larger and more stable.

The evolution of the black hole mass function can be determined by using equation (11) to express $n(M,t)$ as

$$n(M,t) = \begin{cases} n_*\left(\frac{M}{M_*(t)}\right)^{\alpha}\frac{M_{*,i}}{M_*(t)} & M < M_*(t)\ , \\ n_*\left(\frac{M}{M_*(t)}\right)^{\beta}\frac{M_{*,i}}{M_*(t)} & M > M_*(t)\ , \end{cases} \qquad M_*(t) = M_{*,i}\exp\left(\int_{t_i}^{t}\langle\dot{m}(t)\rangle\frac{dt}{t_s}\right)\ , \qquad (27)$$

where $M_{*,i}$ is the mass corresponding to the break in the black hole mass function at a time $t=t_i$. Since we assume that $p(\dot{m})$ is independent of mass, the slopes of the QLF at low and high luminosities match the slopes of $n(M)$, and thus we choose the Boyle et al. (2000) slopes $\alpha = -1.5$ and $\beta = -3.4$ for $n(M)$. We use $M_* = 10^9 M_\odot$ as in §3.2. For $p(\dot{m})$, the values $a = -0.5$, $b = -3.0$, $\dot{m}_{*,i} = 0.9$ then give a reasonable fit to the data at $z = 2$. For our choice of $p(\dot{m}|t)$, the integral for $M_*(t)$ can be done analytically.

Figure 5 uses equations (26) and (27) to determine the black hole mass function and probability function at redshifts $z = 2$, 1, and 0.5. The left hand panels show $M_*(t)$ in the lower windows and $p(\dot{m}|t)$ in the upper windows. The right hand panels show the evolution of the QLF generated by the functions in the corresponding left hand panels, using the methods and assumptions described in §3.2. To reduce the dimensionality of the parameter space, we assume a value of γ_m and then find a value for γ_p that makes the evolved, $z \sim 0.5$ QLF match the Boyle et al. (2000) data at the break luminosity $L_{\rm brk}$. As a characterization of the mean accretion level, we list in the left hand panels the values of the local accretion weighted lifetimes $t_{{\rm acc},z}$ (eq. 25), in units of the Salpeter time t_s, at $z = 2$, 1, and 0.5. These ratios $t_{{\rm acc},z}/t_s$ give the number of e-folds of black hole growth that would occur in a Hubble time $H^{-1}(z)$ if $p(\dot{m})$ stayed constant. In contrast to earlier figures, we plot $\dot{m}^2 p(\dot{m})$ rather than $p(\dot{m})$ itself, because this product gives the contribution to black hole growth (and emissivity of the quasar population) per logarithmic interval of $\dot{m}$. For our adopted forms of $p(\dot{m})$, the largest contribution to growth and emissivity always comes from accretion rates near $\dot{m}_*$.

The upper panels of Figure 5 show an example with $\gamma_m = 0$, so that the normalization of $p(\dot{m}|t)$ evolves but the shape does not. Although we require the predicted QLF to pass through the observed $\Phi(L_{\rm brk})$ at each redshift, the shapes of the model luminosity functions at $z = 1$ and $z = 0.5$ disagree strongly with the data. Generally, evolution of p_* alone cannot reproduce the type of QLF evolution found by Boyle et al. (2000) at low redshift, because the growth of M_* combined with a fixed *relative* distribution of accretion rates shifts the predicted QLF horizontally to a higher break luminosity, while the observed break luminosity declines with time. Reducing p_* reduces the fraction of accreting black holes, but that only produces a vertical drop of the QLF, which cannot fully compensate for the shift to a higher break luminosity.

Since "pure density evolution" models fail to describe the Boyle et al. (2000) data, it is no surprise that our analogous "pure p_* evolution" models also fail. However, Figure 5 shows that black hole growth exacerbates the failures of a model in which only the frequency of fueling activity declines with time, since such models generically predict an increase in the break luminosity with time. We conclude that, if $p(\dot{m})$ is independent of mass, then it must evolve in a way that increases the relative probablilities of low accretion rates in order to make the QLF evolve to lower break luminosities. We will consider an alternative explanation — that $p(\dot{m}|M)$ has a mass dependence that evolves with time — in §4.2.2. However, a decline in typical values of $\dot{m}$ does seem a plausible consequence of decreasing gas fractions and increasing stability of host galaxies, and such changes in galaxy properties are likely to play an important role in producing the observed form of QLF evolution. Furthermore, even if gas fueling rates remain fixed in physical units, they decline in Eddington units as black hole masses increase, thus driving $\dot{m}$ values down if black holes grow

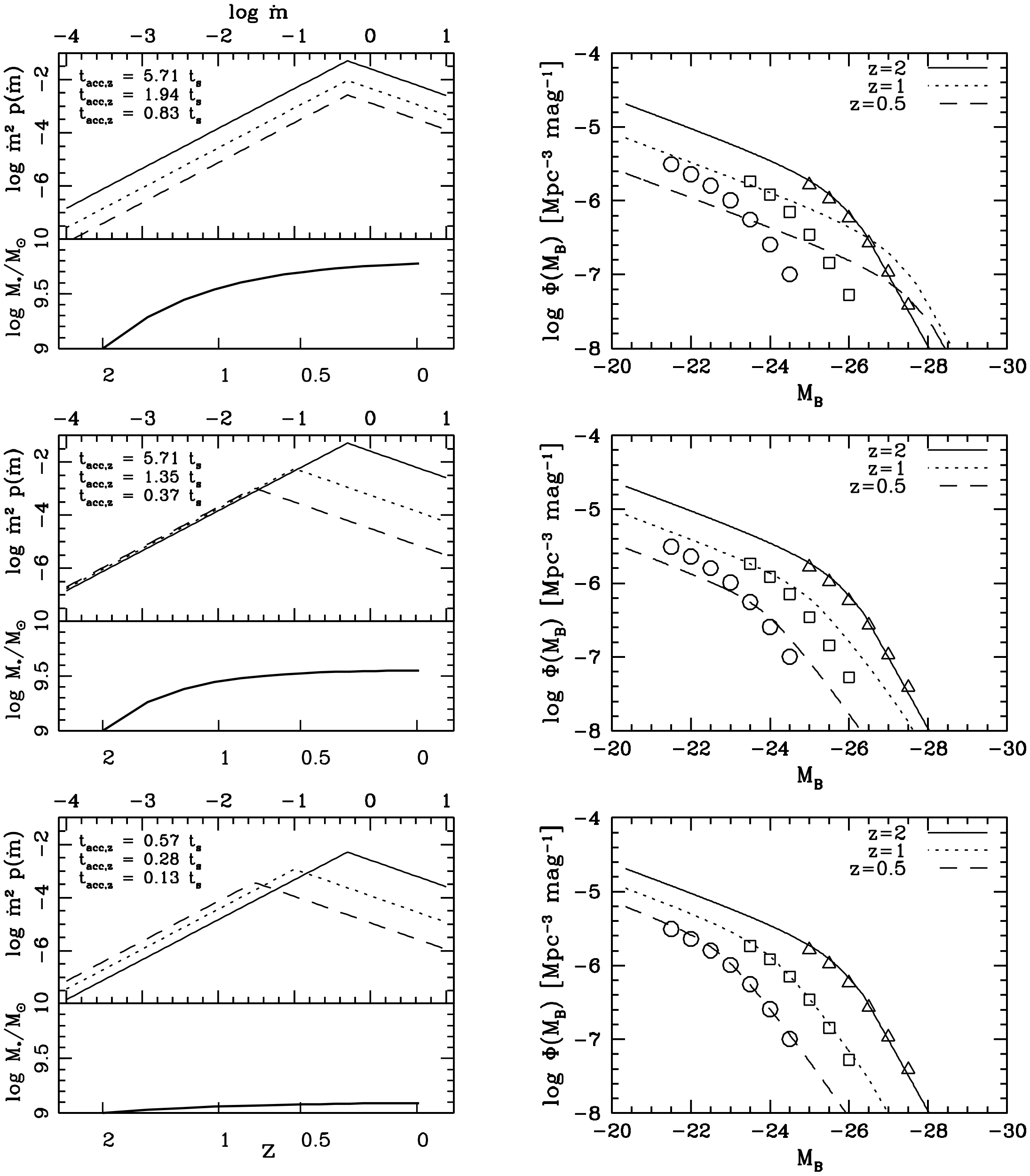

Fig. 5.— Three models of the declining phase of QLF evolution, from $z = 2$ to $z = 0.5$. For each model, left hand panels show $\dot{m}^2 p(\dot{m})$ at $z = 2, 1,$ and 0.5 in the upper window (solid, dotted, and dashed lines, respectively), and $M_*(z)$ in the lower window. Each panel also lists the accretion lifetime $t_{\mathrm{acc},z} = \langle \dot{m} \rangle H^{-1}(z)$ in units of the Salpeter timescale $t_s = 4.5 \times 10^7$ yr, for $z = 2, 1,$ and 0.5 (top to bottom). Right hand panels show the corresponding QLFs at the three redshifts. Points show the fits of Boyle et al. (2000) at $z = 2$ (triangles), $z = 1$ (squares), and $z = 0.5$ (circles) over roughly the absolute magnitude range probed by their data. The top panels show a model in which $\dot{m}_*$ stays fixed and only the normalization p_* declines with time. The middle panels show a model with the same initial $p(\dot{m})$ but declining evolution of $\dot{m}_*$. The bottom panels show a similar model that starts with a higher black hole space density and lower normalization of $p(\dot{m})$, and consequently less black hole growth. Matching the observed evolution of L_{brk} towards lower luminosity requires a decrease in the characteristic accretion rate $\dot{m}_*$, not merely a decrease in duty cycle.

by a significant factor. In their semi-analytic model of the quasar and host galaxy population, Kauffmann & Haehnelt (2000) find that they must account for the decreasing gas supplies and longer dynamical timescales of host galaxies at low redshift, in addition to the decreasing frequency of mergers, to explain the observed evolution of the QLF. In terms of our models, the first two effects are analogous to decreases in $\dot{m}_*$, while the last is analogous to a decrease in p_*.

The models in the middle and bottom rows of Figure 5 incorporate declining $\dot{m}_*$ and match the observed QLF evolution better. The case in the middle row starts with the same $p(\dot{m})$ and $n(M)$ at $z = 2$, but it has $\gamma_m = -2.7$, which moves the break in $p(\dot{m}|t)$ to lower accretion rates as time progresses. Though the match to the data is not perfect, it is much better than before, with the break in $\Phi(L)$ shifting to lower luminosities as the QLF becomes increasingly dominated by the contribution from lower accretion rates. The model in the bottom row starts at $z = 2$ with a black hole space density a factor of ten higher and an accretion duty cycle a factor of ten lower (n_* and p_* increased and decreased by ten, respectively). The QLFs at $z = 2$ are identical because of the exact degeneracy between n_* and p_* discussed in §3.1. However, the reduction in p_* lowers the mean accretion rate $\langle\dot{m}\rangle$, which in turn leads to less black hole growth: this model has $t_{\text{acc},z} < t_s$ at all $z < 2$, and the characteristic mass M_* hardly grows at all. The shift to lower accretion rates coupled with the smaller amount of black hole growth over time yields QLF evolution in good agreement with the data.

Figure 5 shows that the degeneracy between n_* and p_* is broken once evolution is taken into account. If the black hole density is low, then each black hole must accrete more in order to match the observed QLF, and this accretion leads to more rapid evolution of $n(M)$. In the case shown in Figure 5, the model with less black hole growth matches the data better. However, it is possible to start with the initial conditions of the model in the middle panels and match the data nearly as well by dropping $\dot{m}_*$ more rapidly, with $\gamma_m = -3.7$ instead of -2.7. Thus, changes to n_* and p_* can be partly compensated by changes to other parameters. Nonetheless, evolution narrows the range of the $n_* p_*$ degeneracy because models with too much black hole growth (too low n_*) cannot yield a declining break luminosity for any plausible evolution of $p(\dot{m})$. Furthermore, as discussed in §3.3 and further in §5 below, models with different n_* predict different distributions of active black hole masses and accretion rates, different space densities of host galaxies, and of course different underlying $n(M)$, even when they match the same, evolving QLF.

Similar remarks apply to the partial degeneracy between M_* and $\dot{m}_*$ discussed in §3.2. Combinations of M_* and $\dot{m}_*$ that yield similar QLFs at $z \sim 2$ will have more growth of $n(M)$, and thus different evolution, if M_* is lower and $\langle\dot{m}\rangle$ consequently higher.

4.2.2. Declining Phase with Mass-Dependent $p(\dot{m})$

As shown in §3.4, mass dependence of $p(\dot{m})$ can break the link between the shape of the black hole mass function and the shape of the luminosity function, adding considerable freedom to models of the QLF. In an evolutionary calculation, one must also account for the influence of mass-dependent growth on the shape of the black hole mass function. If the more numerous, low mass black holes have a higher probability of being active, then they grow faster than the high mass black holes, and the mass function steepens. Conversely, if high mass black holes are more active, then they grow faster and the mass function becomes shallower with time. It is intuitively useful to think of this behavior in graphical terms. On a log-log plot, a mass-independent $p(\dot{m})$ causes the mass function to shift horizontally in a coherent fashion, maintaining its shape. Faster growth of low mass black holes allows the low end of $n(M)$ to translate faster and "catch up" with the high end. Conversely, faster growth of high mass black holes allows the high end of $n(M)$ to stretch away from the low end. As always (eq. 10), it is only the mean accretion rate $\langle\dot{m}(M)\rangle$ that matters for determining the evolution of $n(M)$, and one can calculate the evolution exactly by assuming that all black holes of a given mass accrete at this rate.

To describe the evolution of $n(M)$ mathematically, we define $g(M,t) = M_i/M$, where M_i and M represent black hole masses at times t_i and t, respectively. Matching number densities in the equivalent mass intervals at the two times then implies that $n(M)dM = n_i(M_i)dM_i$, and thus

$$n(M) = n_i(M_i)\frac{dM_i}{dM} = n_i(M_i)\left[g(M,t) + M\frac{\partial g}{\partial M}\right] \ . \qquad (28)$$

If $g(M,t)$ is independent of mass, then we have simple remapping of $M_i \rightarrow M$ and renormalization by $g(M,t)$, recovering the result (11). For mass-dependent $\langle\dot{m}\rangle$, the value of $n(M)$ at mass M may be higher or lower than the mass-independent case depending on the sign of $\partial g/\partial M$. Equation (28) allows the numerical calculation of $n(M)$ for any specified $p(\dot{m}|M,z)$, since this determines $g(M,t)$. However, the exponential relation between the growth factor and $\int\langle\dot{m}\rangle dt$ generally means that any simple analytic form of $n(M)$ is lost once black holes grow significantly.

We have previously adopted mass-independent $p(\dot{m})$ as a mathematical convenience, but consideration of black hole growth suggests that this choice is not completely arbitrary. Suppose that the black holes in some mass range have a high $\langle\dot{m}\rangle$ relative to their peers because they tend to reside in galaxy hosts that can feed them more efficiently. These black holes e-fold in mass more rapidly than others, and their fueling rates *in Eddington units* therefore drop more quickly (or grow more slowly), bringing them back into line. This regulating mechanism suppresses mass-dependence of $\langle\dot{m}\rangle$, causing $n(M)$ to change shape until it approaches the "fixed point" solution of mass-independent $\langle\dot{m}\rangle$, after which it evolves in a self-similar fashion. Constancy of $\langle\dot{m}\rangle$ does not necessarily imply constancy of $p(\dot{m})$, but this regulation argument suggests that $p(\dot{m})$ might be approximately mass-independent in an average sense at redshifts near the peak of quasar activity.

The regulating mechanism may lose force at low redshift, when black holes no longer grow by substantial factors and gas merely trickles onto full grown systems. Thus, we might expect stronger mass dependence of $p(\dot{m})$ in the declining phase of quasar evolution. We have shown in §4.2.1 that matching the observed shift of $L_{\rm brk}$ to lower luminosity requires a decline in characteristic $\dot{m}$ values if $p(\dot{m})$ is independent of mass. However, mass-dependent $p(\dot{m})$ offers another possibility: activity could decline preferentially in more massive black holes between $z = 2$ and $z = 0$, thus driving the break luminosity down as the typical mass of *active* systems declines.

To create a model along these lines, it is helpful first to consider the case where there is no evolution of $n(M)$ at all, so that one can infer the required $p(\dot{m}|M, z)$ from a simple graphical argument relating the vertical shift of $\Phi(L)$ to the horizontal shift of $L_{\rm brk}$. We consider a double power-law QLF with slopes α and β below and above $L_{\rm brk}$ respectively, and we assume pure luminosity evolution with $\Phi(L_{\rm brk}) =$ constant and $L_{\rm brk}(t_2) < L_{\rm brk}(t_1)$ for $t_2 > t_1$. For maximum contrast with the model in §4.2.1, we assume that the mass dependence of $p(\dot{m})$ enters only in the normalization $p_*(M)$, not in the slopes or in the characteristic accretion rate $\dot{m}_*$. In other words, the relative distribution of accretion rates remains the same at all times for all active black holes, but the probability of a black hole being active at all depends on its mass, in a redshift-dependent manner. Relating the amplitudes of $\Phi(L)$ at times t_1 and t_2 to the horizontal shift of $L_{\rm brk}$ from time t_1 to t_2 then implies

$$\frac{\Phi(L,t_2)}{\Phi(L,t_1)} = \begin{cases} k^{\alpha} & L < L_{\rm brk}(t_2) \\ k^{\beta} & L > L_{\rm brk}(t_1) \end{cases}, \qquad k = \frac{L_{\rm brk}(t_2)}{L_{\rm brk}(t_1)}\,, \tag{29}$$

with a more complicated dependence in the range $L_{\rm brk}(t_2) < L < L_{\rm brk}(t_1)$. Since $n(M)$ is constant, any shift in the QLF must be produced by a shift in $p_*(M)$, and thus,

$$\frac{p_*(M,t_2)}{p_*(M,t_1)} = \begin{cases} k^{\alpha} & M < M_{\rm min} \\ k^{\beta} & M > M_{\rm max} \\ k^{\alpha} + (k^{\beta} - k^{\alpha})U(M) & M_{\rm min} < M < M_{\rm max}\,, \end{cases} \tag{30}$$

where $M_{\rm min}$ is the minimum mass of a black hole that can generate $L = L_{\rm brk}(t_2)$ and $M_{\rm max}$ is the maximum mass of a black hole that can generate $L = L_{\rm brk}(t_1)$. Here $U(M)$ is some function that goes from zero to one as mass goes from $M_{\rm min}$ to $M_{\rm max}$, and it can be tuned to reproduce the observed $\Phi(L)$ in the break region. With equation (30) as a starting guess, we can find by numerical iteration a solution with mass-dependent $p_*(M)$ that reproduces the observed evolution and self-consistently incorporates the implied growth in $n(M)$.

Figure 6 shows a model in which we assume that $p(\dot{m})$ is independent of mass at $z = 2$ and that subsequent evolution of the QLF (left panel) is driven by the mass dependence of $p(\dot{m}|M)$ (right panel). We choose a normalization of the black hole mass function, $n_* M_* = 1.2 \times 10^{-4}$ Mpc^{-3}, that corresponds to a fairly short quasar lifetime, $t_{\rm acc} = 7 \times 10^6$ yr (from $z = 2$ to $z = 0$), so that the mass-dependent growth does not severely distort the double power-law form of $n(M)$ that our initial guess at $p(\dot{m}|M)$ assumes. This model fits the Boyle et al. (2000) data as well as the model with mass-independent $p(\dot{m})$ shown in the bottom panel of Figure 5. With regard to $\Phi(L)$ alone, the two models are effectively degenerate, but $L_{\rm brk}$ in the mass-dependent model evolves to lower luminosities because high mass black holes are less likely to be active at low redshift. Thus, relative to the mass-independent model, this model predicts that luminous AGN at low redshift consist primarily of low mass black holes with high $L/L_{\rm Edd}$, and it predicts a narrower range of $\dot{m}$

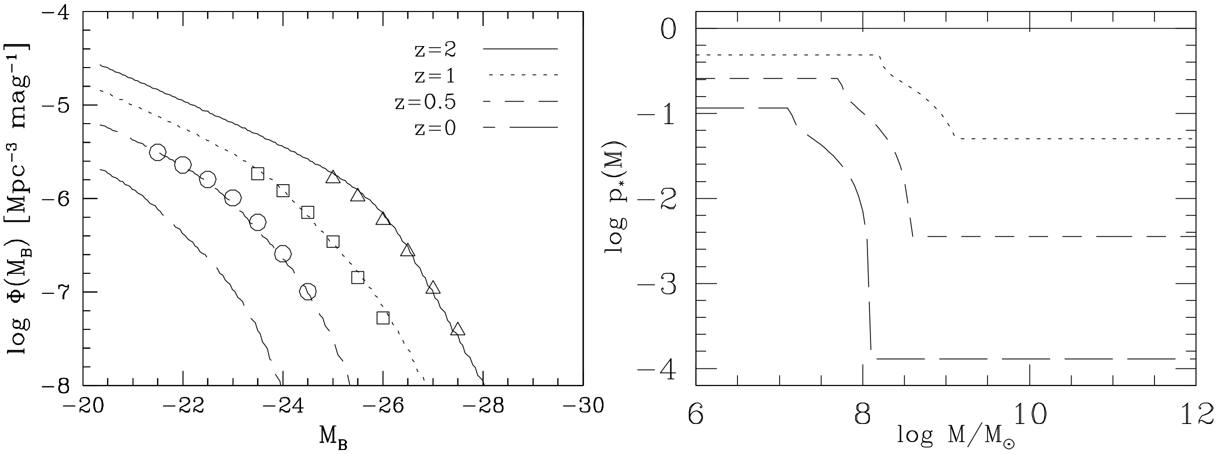

Fig. 6.— An alternative model for the declining evolution of the QLF, in which the characteristic accretion rate $\dot{m}_*$ remains constant but the mass dependence of $p(\dot{m}|M)$ evolves to suppress activity preferentially in high mass black holes at lower redshifts. As in Figure 5, triangles, squares, and circles in the left hand panel represent the Boyle et al. (2000) observational fits at $z = 2$, 1, and 0.5, respectively, while lines show the model predictions (including an extrapolation to $z = 0$). The right hand panel shows the evolving mass dependence of the normalization $p_*(M)$, which is proportional to the duty cycle of black holes of mass M. The model assumes our usual double power law forms of $n(M)$ and $p(\dot{m})$, with $M_* = 10^9 M_\odot$ and a mass-independent $p(\dot{m})$ at $z = 2$, and $\dot{m}_* = 0.5$ at all redshifts.

values. We compare the two models' predictions for the mass distribution of active systems in §5 below (Fig. 16).

4.2.3. Growing Phase

We now turn to the redshift interval $z \sim 5$ to $z \sim 2$, during which $\Phi(L)$ first grows rapidly, then reaches a plateau between $z \sim 3$ and $z \sim 2$ (see, e.g., Warren, Hewett, & Osmer 1994; Schmidt, Schneider, & Gunn 1995; Pei 1995). In the range $z = 3.6 - 5.0$, Fan et al. (2001) provide the best measurements of the bright end of the luminosity function, while Wolf et al. (2003) give constraints at fainter luminosities. To cover the gap between $z = 3.6$ and the Boyle et al. (2000) measurements at $z = 2$, we use the measurements of Warren, Hewett, & Osmer (1994), with a median redshift $z \approx 3.25$. We use the $\Omega_m = 1$ cosmological model, since all of these papers give results for this case, and we adopt $h = 0.5$ so that the age of the universe is realistic. We convert the Warren, Hewett, & Osmer (1994) space densities and absolute magnitudes from $h = 0.75$ to $h = 0.5$, and we convert their AB(λ1216Å) magnitudes to B-band magnitudes using M_B =AB(λ1216Å)-0.605, assuming $f_\nu \propto \nu^{-0.5}$ for the conversion of AB(λ4400Å) to AB(λ1216Å). Our goal here is not to model the data in detail but merely to illustrate what kinds of $p(\dot{m})$ and accompanying $n(M)$ evolution can fit the general trends.

For simplicity, we restrict ourselves to the class of models in which $p(\dot{m})$ is a double power-law independent of mass and the characteristic accretion rate $\dot{m}_*$ is constant from $z = 5$ to $z = 2$. At $z = 2$, we assume a double power-law $n(M)$ with $M_* = 10^9 M_\odot$ and slopes $\alpha = -1.5$ and $\beta = -3.4$. The evolution of the QLF is then determined by the evolution of the amplitude of $p(\dot{m})$, which we assume has a piecewise power-law form, $p_*(t) \propto t^{\gamma_p}$ in the interval between each of the redshifts where we match the QLF, though we allow γ_p to be different from one interval to another. Our general procedure is to take a model that fits the QLF data at $z = 2$, then evolve it to higher redshift, iteratively solving for values of γ_p in each redshift interval so as to match the observed amplitude of the QLF at $M_B \sim -26.5$, using the Warren, Hewett, & Osmer (1994) data at $z = 3.25$ and the Fan et al. (2001) data at $z = 3.75, 4.15$, and 4.7. Iteration is necessary because we must calculate the time integrated mean accretion rate for the given γ_p and reduce black hole masses by the corresponding factor, before calculating the QLF with the evolved $p(\dot{m})$. Because mass-independent $p(\dot{m})$ leads to self-similar evolution of $n(M)$ and identical asymptotic slopes for $n(M)$ and $\Phi(L)$ (see §3.1), this restricted class of models cannot explain the apparent change in the bright-end slope of the QLF (see Fan et al. 2001) between $z \sim 4$ and $z \sim 2$.

Figure 7 shows two qualitatively different solutions that reproduce the observed evolution of the QLF amplitude. The first of these solutions, shown in the top panels, has a relatively high normalization of the black hole mass function at $z = 2$, with space density $n_* M_* = 1.2 \times 10^{-4}$ Mpc^{-3}. The high space density leads to a low normalization of $p(\dot{m})$ at each redshift, and consequently to little growth of black hole masses; the lifetimes $t_{\mathrm{acc},z}$ are shorter than t_s at all redshifts, and the value of M_* increases only slightly over the entire range $z = 4.7$ to $z = 0$. (Note

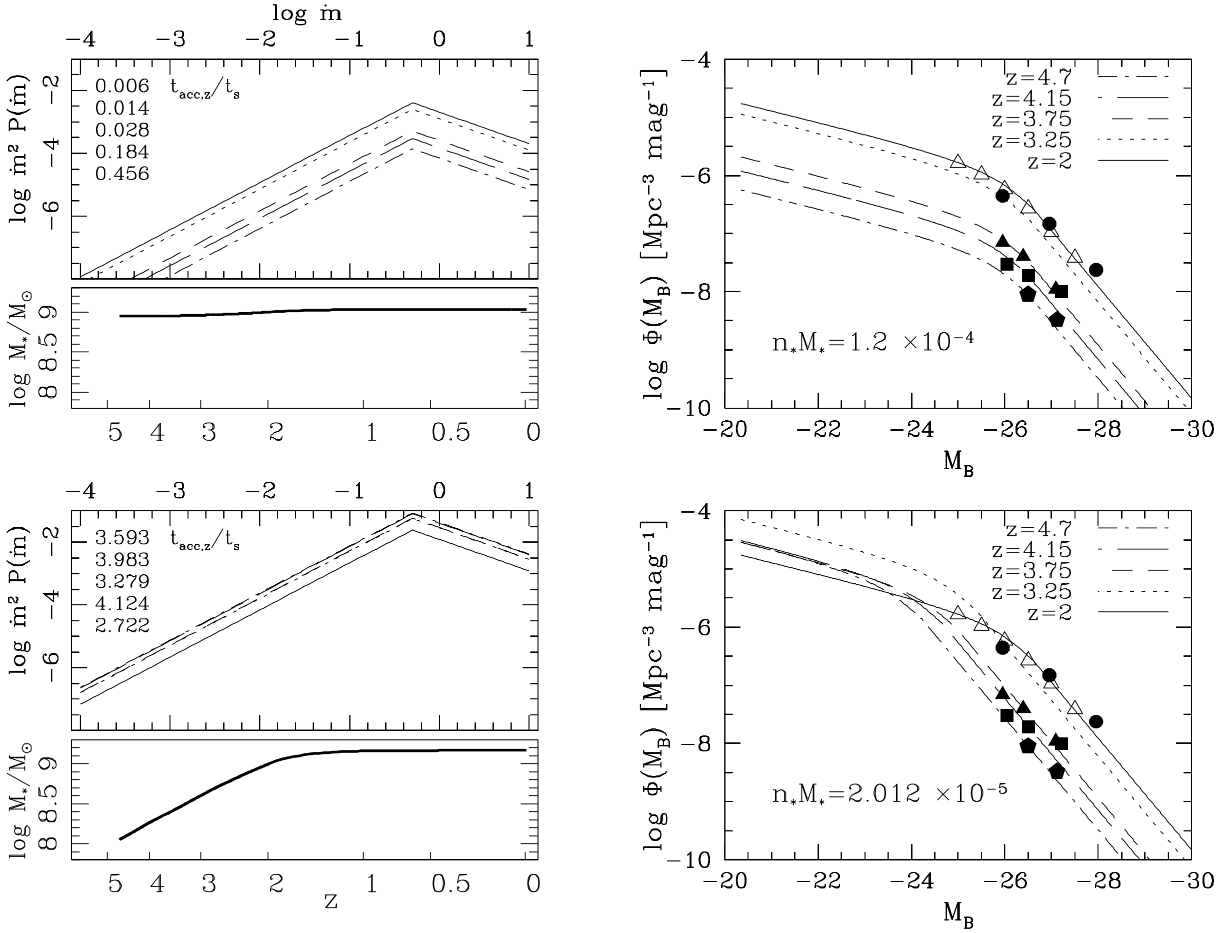

Fig. 7.— Like Figure 5, but for the growing phase of QLF evolution from $z \sim 5$ to $z \sim 2$. Points in the right hand panels are based on Fan et al. (2001) at $z \sim 4.7$, 4.15, and 3.75 (pentagons, squares, filled triangles), on Warren, Hewett, & Osmer (1994) at $z \sim 3.25$ (circles), and on Boyle et al. (2000) at $z \sim 2$ (open triangles). Model parameters are chosen to reproduce the amplitude of the observed QLF at $M_B \sim -26.5$. The top row shows a model with a high black hole space density at $z = 2$, $n_* M_* = 1.2 \times 10^{-4}$ Mpc^{-3}, in which case there is little growth of black hole masses between $z = 5$ and $z = 2$ and the QLF evolution is driven by increasing $p(\dot{m})$. The bottom row shows a model with a lower black hole space density and, therefore, more accretion per black hole. In this model, the growth of the QLF is driven mainly by the increasing masses of black holes.

that we show the low redshift evolution of M_* inferred by continuing the model past $z = 2$ using the Boyle et al. [2000] data as in the previous section.) The growth in the amplitude of the QLF is therefore driven by the steadily rising amplitude of $p(\dot{m})$, with no significant contribution from growth in $n(M)$ itself.

While this high space density solution fits the QLF data reasonably well by construction, it seems rather implausible on physical grounds. The minimal evolution of M_* means that all of the observed quasar activity represents a negligible contribution to the growth of black holes. Instead, all of the supermassive black holes must have been assembled in some unseen manner before $z = 5$, and the observed evolution of the QLF reflects a gradual "turning on" of these black holes at lower redshift. In such a model, the "seed" black holes formed at $z > 5$ are essentially the same black holes that are present today, though mergers may have converted many low mass black holes into a smaller number of high mass systems.

The bottom panels show a solution near the opposite extreme, with a lower space density $n_*M_* = 2.012 \times 10^{-5}$ Mpc^{-3} at $z = 2$. Matching the QLF now requires a higher normalization of $p(\dot{m})$ at each redshift, and the accretion weighted lifetimes are in the range $t_{\mathrm{acc},z} \sim 0.4 - 2.7t_s$. There is roughly a factor of ten growth in M_* between $z = 5$ and $z = 2$, and this growth plays a central role in the evolution of the QLF. In contrast to the high space density case, the amplitude of $p(\dot{m})$ is approximately constant from $z = 5$ to $z = 3$, with a slight drop to $z = 2$. In this solution, therefore, the observed quasar activity traces the growth of the black hole population from much smaller seeds present at $z = 5$, with a roughly constant distribution of accretion rates in Eddington units during the growing phase of quasar evolution. The break in the QLF moves to higher luminosity as M_* grows, so measurements probing to lower luminosities at high redshift could distinguish this solution from the short lifetime solution shown in the top panels.

If the $n(M)$ normalization is reduced just slightly further, below $n_*M_* \approx 2.01 \times 10^{-5}$ Mpc^{-3}, then the solution for the evolving $n(M)$ requires unphysical, negative densities at $z = 5$. Thus, within our assumptions, there is a minimum allowed space density of black holes, which corresponds to the limit in which the observed QLF traces all of the accretion onto the black hole population. We can see why this is so by recalling Soltan's (1982) argument that the integrated bolometric emissivity of the quasar population determines, for an assumed efficiency, the mass density $\rho_{\mathrm{bh}} = \int_0^\infty Mn(M)dM$ of the black hole population. We have adopted a form for $n(M)$, and the evolutionary calculation allows us to, in effect, correct the observed emissivity for the contribution from lower luminosity systems. Since we have also specified the efficiency $\epsilon(\dot{m})$ and the bolometric correction L_{bol}/L_B, matching the observational data points determines ρ_{bh} at $z = 2$ and thereby fixes the normalization of $n(M)$. Our results thus show that a suitable extension of the Soltan (1982) argument, aided by some auxiliary assumptions and a measurement of the QLF at $z \sim 2$, can predict the black hole mass function $n(M)$ itself, not just the integral ρ_{bh}. We will explore this idea in future work, with attention to the sensitivity of the predictions to the auxiliary assumptions and to uncertainties in the observational data.

4.3. Mergers

The general equation (12) for the evolution of the black hole mass function has a very limited set of analytic solutions, even if one considers only the merger terms and ignores accretion driven growth. Realistic calculations incorporating merger driven growth will probably need to be done numerically, with some *a priori* model (based on galaxy merger trees, for example) for what merger rates should be. Here we will investigate some simple, analytically solvable cases that can provide insight into the generic effects of mergers on the black hole mass function.

First, we consider a binary merger model in which there is a probability f per unit time that a given black hole merges with another black hole of equal mass to form a new black hole of twice the original mass. In the absence of accretion, a counting argument yields the equation governing $n(M,t)$,

$$\frac{\partial n(M,t)}{\partial t} = -fn(M,t) + \frac{1}{4} fn\left(\frac{M}{2},t\right). \tag{31}$$

The first term on the r.h.s. is the sink representing loss of black holes of mass M to mergers, and the second is the source representing creation by mergers of systems with mass $M/2$, with a factor of 1/4 to account for the replacement of two black holes by one and the factor of two growth in the dM interval. The only solution to (31) in which $n(M)$ maintains its shape over time is a pure power-law of the form

$$n(M,t) = n_*(t)\left(\frac{M}{M_*}\right)^{\alpha}, \tag{32}$$

where we have included all the time dependence in n_* because the effects of n_* and M_* are degenerate for a pure power-law. With this form, the solution to (31) is

$$n_*(t) = n_*(t_i)\exp\left[\int_{t_i}^{t}\left(2^{-(\alpha+2)} - 1\right) f(t)dt\right]. \tag{33}$$

If $f(t)$ is constant, then $n_*(t)$ evolves exponentially in time, while $f(t) \propto t^{-1}$ yields $n_*(t)$ evolving as a power-law with slope of $\left(2^{-(\alpha+2)} - 1\right) t_i f(t_i)$.

The important general feature of the solution (33) is that the *sign* of the evolution depends on the slope α of the mass function. For the critical value $\alpha = -2$, mergers do not change $n(M)$ at all, because the source and sink terms in equation (31) balance. If $\alpha < -2$ then $n(M)$ increases with time, and if $\alpha > -2$ then $n(M)$ decreases with time. For a steep $n(M)$, the black holes added to a given range of mass from mergers of lower mass objects exceed the number lost to higher masses. For a shallow $n(M)$, on the other hand, mergers consume more black holes in a given mass range than they create. While the solution (33) is specific to our restricted model, different behavior for steep and shallow slopes of $n(M)$ follows from mass conservation, so we expect it to hold quite generally.

The pure power-law $n(M)$ adopted above cannot hold for all masses because the implied total mass density of black holes would be infinite. However, the behavior of the single power-law solution gives insight into the more general case of a mass function that changes slope from low to high masses. For example, a double power-law has a steep high mass end where mergers drive $n(M)$ up with time and a shallow low mass end where mergers drive $n(M)$ down. Over the range in mass near the break, the shape changes as the number of high mass black holes grows and low mass black holes decreases, making the break smoother and shifting it to lower masses. Again, we expect these effects of mergers to be fairly generic.

For a pure power-law $n(M)$, we can also include the accretion term of equation (12) in the calculation, obtaining the solution

$$\frac{\dot{n}}{n} = \frac{\dot{n}_*}{n_*} = -\frac{\langle \dot{m}(t) \rangle}{t_s}(1+\alpha) + \left(2^{-(\alpha+2)} - 1\right) f(t) \ . \tag{34}$$

The sign of the accretion term also depends on α, but here the critical slope is -1 instead of -2, reflecting the fact that accretion adds mass to the black hole population while mergers do not. Equation (34) can be integrated analytically if accretion and merger rates have same time dependence, $\langle \dot{m}(t) \rangle = \langle \dot{m}(t_i) \rangle h(t)$ and $f(t) = f(t_i)h(t)$, with $h(t)$ an arbitrary function having $h(t_i) = 1$. The solution is

$$n_*(t) = n_*(t_i) \exp\left(\left[-\frac{\langle \dot{m}(t_i) \rangle}{t_s}(1+\alpha) + \left(2^{-(\alpha+2)} - 1\right) f(t_i)\right] \int_{t_i}^{t} h(t)dt\right) . \tag{35}$$

Up to factors that are typically of order unity, the relative importance of accretion and mergers depends on the value of $\langle \dot{m}(t_i) \rangle$ relative to $t_s f(t)$, the average number of mergers per Salpeter time. However, the pre-factors can greatly diminish one term or the other close to the critical slopes $\alpha = -1$ or $\alpha = -2$, and for $-1 > \alpha > -2$ accretion and mergers affect $n(M)$ in opposite directions.

For our illustrative model calculations in §5, we do not want to assume a pure power-law $n(M)$, and we will therefore adopt another very simple prescription for mergers, similar to that of Richstone et al. (1998). We assume that accretion has negligible impact on $n(M)$ after some time t_1, which is true if $\langle \dot{m}(t_1) \rangle t_1 \ll t_s$ and $\langle \dot{m}(t) \rangle$ falls as t^{-1} or faster. At some later time, t_2, every black hole with initial mass M_1 is assumed to have merged with $(f_m - 1)$ other black holes of mass M_1 to make one black hole of mass $M_2 = f_m M_1$. The black hole mass functions at t_1 and t_2 are related by the transformation

$$n_2(M_2)dM_2 = \frac{1}{f_m} n_1\left(M_1 = \frac{M_2}{f_m}\right) \frac{dM_1}{f_m}, \tag{36}$$

or simply

$$n_2(M) = f_m^{-2} n_1(M/f_m). \tag{37}$$

In this model as in our previous model, mergers drive $n(M)$ up when the logarithmic slope is $\alpha < -2$ and down when $\alpha > -2$. If $n(M)$ is a double power-law at time t_1, then at time t_2 it is still a double power-law, with M_* larger by a factor f_m and the normalization lower by a factor f_m^2.

In both of our merger models, we assume that mergers conserve the total mass of the black hole population, and simply redistribute it from low mass systems to high mass systems. However, it is also possible for mergers to *decrease* the total mass of the population, at least those black holes that reside at the centers of galaxies and have the potential to become quasars. This can happen if multiple mergers produce triple or quadruple systems that lead to ejection of one or more of the black holes from the galaxy (e.g., Valtonen et al. 1994). Even a merging binary system can potentially be ejected by a gravitational radiation "rocket" effect, though this seems more likely to be important in shallow potential wells hosting low mass black holes (see Redmount & Rees 1989; Madau et al. 2003). Finally, a binary could remain at the galaxy center but radiate a significant fraction of the mass of its progenitors in gravity waves during the merger event (Yu & Tremaine 2002 and references therein). We will not consider any of these possibilities in detail here, but we note that in all these cases the critical slope at which $n(M)$ grows rather than declines would be steeper than -2, since a given bin would lose mass to mergers at the same rate as before but would gain mass at a lower rate.

5. Illustrative Scenarios

We now utilize the framework developed above to construct models that illustrate different plausible scenarios for the evolution of the quasar population. The simplest scenario is that the luminous quasar population is dominated by black holes accreting with thin-disk efficiency $\epsilon_{0.1} \approx 1$, all radiating with a "standard" SED (e.g., Elvis et al. 1994), and that the growth of black holes is driven by the observed accretion. The key parameter of this scenario is the typical quasar lifetime, which is linked in turn to the space density of black holes. However, there are many potential variations on this theme, including the possibility that a large fraction of quasar activity is obscured, that black hole growth is substantially affected by low redshift mergers, or that substantial black hole growth occurs though low efficiency ADAF accretion. Our models here illustrate each of these physically distinct possibilities, including long and short quasar lifetimes for the simplest scenario. For simplicity, we will consider only models with $p(\dot{m}|z)$ independent of mass. We examine only the regime from $z = 2$ to $z = 0$ and choose parameters so that each model approximately reproduces the observed optical QLF. Our goal is to determine what other observables, such as the QLF in other bands, the masses of active black holes at different luminosities, the black hole mass function itself, and the space density of host systems are most likely to discriminate among these scenarios. We include some comparisons to recent estimates of these observables, but our emphasis is mainly on the differences among the models themselves, since we have not made any adjustments to model parameters to try to match these other data.

5.1. Model Parameters

The parameters of the five models are summarized in Table 2. Our baseline parameters are similar to those used in the evolutionary calculations of §4.2.1. We adopt a double power-law $n(M)$ with slopes $\alpha = -1.5$ and $\beta = -3.4$, which are required to match the asymptotic slopes of the Boyle et al. (2000) QLF for cases where $p(\dot{m})$ is independent of mass. We adopt $M_* \approx 10^9 M_\odot$, making the Eddington luminosity lM_* close to the observed break luminosity at $z = 2$. Except for the ADAF model (discussed below), we adopt a double power-law $p(\dot{m})$ with parameters $a = -0.5$ and $b = -3.0$. We start with a characteristic accretion rate $\dot{m}_* = 0.5$ and evolve it to lower redshift as $\dot{m}_* \propto t^{\gamma_m}$. The normalization p_* evolves as t^{γ_p}, with different γ_p values from $z = 2$ to $z = 1$, $z = 1$ to 0.5, and $z = 0.5$ to 0. The values of γ_m for each model are chosen to give a reasonable match to the observed evolution of $L_{\rm brk}$ given the model's predicted growth of $n(M)$, and the values of γ_p are then chosen by matching the observed amplitude of the QLF. Table 3 summarizes the quantitative evolution of the five models, giving the values of M_*, the comoving black hole mass density $\rho_{\rm bh}$, and the mean accretion rate $\langle\dot{m}\rangle$ at redshifts 2, 1, 0.5, and 0, and the accretion weighted lifetime $t_{\rm acc} = \int_2^z \langle\dot{m}(t)\rangle dt$ at $z = 1$, 0.5, and 0.

For our short quasar lifetime (short-t_q) model, we normalize the black hole mass function at $z = 2$ to $n_* M_* = 1.2 \times 10^{-4}$ Mpc^{-3}. Because of the high space density, the normalization of $p(\dot{m})$ is relatively low, and the accretion weighted lifetime between $z = 2$ and $z = 0$ is $t_{\rm acc} = 9.5 \times 10^6$ yr, which implies little growth of black hole masses over this redshift range (see §4.1). For the long-t_q model, we reduce the black hole space density at $z = 2$ by a factor of ~ 6, to $n_* M_* = 2.012 \times 10^{-5}$ Mpc^{-3}, thus continuing the growing phase model illustrated in the lower panels of Figure 7. The corresponding value of $t_{\rm acc}$ is 5.3×10^7 yr, implying about one e-fold of mass growth from $z = 2$ to $z = 0$ (compared to an order of magnitude growth from $z = 5$ to $z = 2$, as shown in Figure 7). The factor of six difference in lifetimes is small compared to the full range of quasar lifetimes discussed in the recent literature (Martini 2003 and references therein), but the two models straddle the boundary between significant low-z accretion growth and minimal low-z accretion growth. Solid and dotted curves in Figure 8 show the B-band QLFs predicted by the short- and long-t_q models, respectively. By construction, they match the Boyle et al. (2000) data well at $z = 2$, 1, and 0.5.

The $z = 0$ data in Figure 8 come from Wisotzki (2000), based on the Hamburg/ESO quasar survey. They lie close to an extrapolation of the Boyle et al. (2000) evolution model to $z = 0$. We have chosen γ_p values so that each model passes through these data at $M_B \approx -22$, but we have not attempted to reproduce the shape, so the model predictions do not nearly overlap as they do at higher redshift. The short-t_q model agrees well with the Wisotzki (2000) data. The long-t_q model predicts a higher amplitude of the QLF at high luminosities, mainly because greater black hole growth gives it a higher value of M_* at $z = 0$, and its luminosity function is steeper than the data. The discrepancy could be reduced if we allowed $\dot{m}_*$ to drop more rapidly between $z = 0.5$ and 0, instead of extrapolating the behavior that fits from $z = 2$ to $z = 0.5$.

For the obscured model, we essentially take the short-t_q model and multiply p_* by five, assigning 20% of the active systems at each redshift the standard Elvis et al. (1994) SED and the remaining 80% the obscured SED of Table 1. The obscured model's optical luminosity function is shown by the dashed lines in Figure 8, and it also closely matches the observed evolution. A steepening of the evolution of $p(\dot{m})$ to lower accretion rates is required to balance the increased black hole

Table 2. Input Model Parameters

	Short-t_q	Long-t_q	Obscured	Merger	ADAF
$M_*(z=2)$	1	1	1	1	1.25
$n_* M_*(z=2)$	1.2×10^{-4}	2.012×10^{-5}	1.2×10^{-4}	1.2×10^{-4}	3.0×10^{-5}
$\dot{m}_*(z=2)$	0.5	0.5	0.5	0.5	0.5
$p_*(z=2)$	0.016	0.097	0.08	0.016	0.043
$\gamma_m(z=2-0)$	-2.7	-3.9	-3.2	-3.9	-3.1
$\gamma_p(z=2-1)$	2.84	4.51	3.07	5.33	1.60
$\gamma_p(z=1-0.5)$	2.72	5.74	3.67	5.71	2.79
$\gamma_p(z=0.5-0)$	3.75	8.11	5.32	8.83	3.50

Note. — Defining parameters of the five illustrative models discussed in §5. All models assume a double power-law $n(M)$ with $\alpha = -1.5$, $\beta = -3.4$, and a mass-independent, double power-law $p(\dot{m})$ with $a = -0.5$, $b = -3$, except that the ADAF model has $a = -1.5$ and a factor of 16 boost to $p(\dot{m})$ in the range $10^{-4} \leq \dot{m} \leq 10^{-2}$. Models start at $z = 2$ with the tabulated values of M_* (in $10^9 M_\odot$), $n_* M_*$ (in comoving Mpc^{-3} for $\Omega_m = 1$, $h = 0.5$), $\dot{m}_*$, and p_*. Over the indicated redshift intervals, $\dot{m}_*$ evolves as t^{γ_m} and p_* as t^{γ_p}. The obscured model has a 4:1 ratio of obscured to unobscured systems. The merger model incorporates a factor $f_m = 1.8$ growth by equal mass mergers in each redshift interval $z = 2-1$, $1-0.5$, and $0.5-0$. Note that in all five models the amplitude of $p(\dot{m})$ drops with time at all $\dot{m}$ even though the γ_p values are positive, since $\dot{m}_*$ decreases rapidly with time.

Table 3. Evolutionary Values

	Short-t_q	Long-t_q	Obscured	Merger	ADAF
$M_*/10^9 M_\odot$					
$z = 2$	1	1	1	1	1.25
$z = 1$	1.14	1.95	1.73	2.06	4.23
$z = 0.5$	1.19	2.36	1.98	3.92	5.65
$z = 0$	1.24	3.28	2.25	7.92	7.40
$\rho_{\rm bh}/10^5 M_\odot {\rm Mpc}^{-3}$					
$z = 2$	3.26	0.55	3.26	3.26	1.02
$z = 1$	3.72	1.07	5.63	3.74	3.44
$z = 0.5$	3.88	1.29	6.45	3.94	4.61
$z = 0$	4.02	1.79	7.34	4.42	6.03
$\langle \dot{m} \rangle$					
$z = 2$	0.0054	0.032	0.027	0.0054	0.071
$z = 1$	0.0014	0.0054	0.0043	0.0015	0.0084
$z = 0.5$	4.4×10^{-4}	0.0022	0.0014	6.1×10^{-4}	0.0033
$z = 0$	1.6×10^{-4}	0.0027	7.1×10^{-4}	0.0011	0.0015
$t_{\rm acc}/10^7 {\rm yr}$					
$z = 2 - 1$	0.59	3.01	2.46	0.62	5.45
$z = 2 - 0.5$	0.79	3.86	3.08	0.86	6.79
$z = 2 - 0$	0.95	5.35	3.66	1.38	8.00

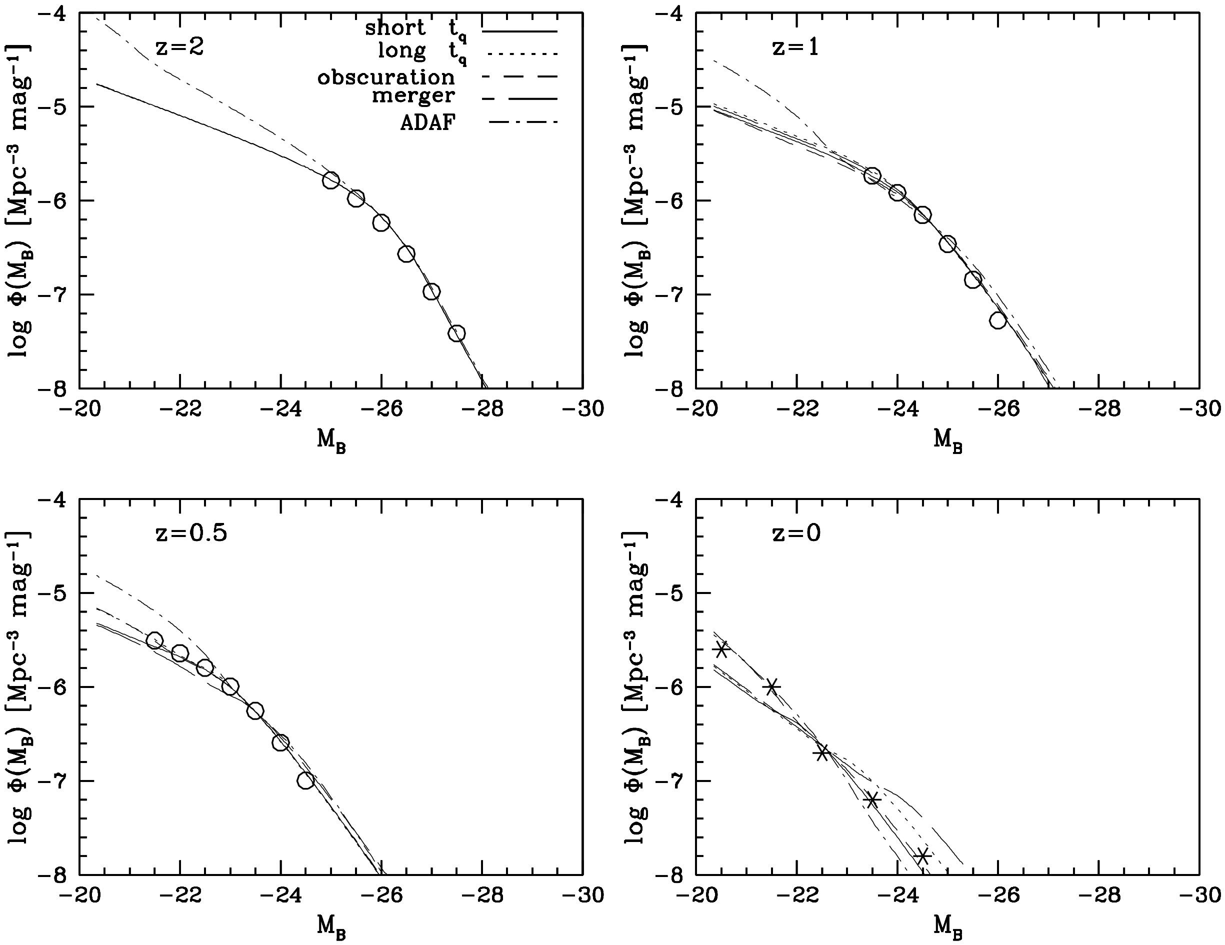

Fig. 8.— Evolution of the B-band QLF for the five illustrative models discussed in §5. In each panel, solid and dotted lines show results for two models dominated by thin-disk accretion, with short and long quasar lifetimes, respectively. Short-dashed lines show a model with a large fraction of obscured quasars, long-dashed lines a model in which mergers contribute substantially to the evolution of the black hole mass function, and dot-dashed lines a model with high probability of low $\dot{m}$ accretion, leading to significant black hole growth in an ADAF mode. Open circles in the $z = 2$, 1, and 0.5 panels show the Boyle et al. (2000) QLF fit over the observed range of luminosities at the indicated redshift. Asterisks in the $z = 0$ panel show the QLF estimate of Wisotzki (2000).

growth from obscured accretion. Note that we could also have implemented the obscured scenario by increasing the $n(M)$ normalization n_* by a factor of five at $z = 2$ and keeping p_* the same, but then $\rho_{\rm bh}(z = 0)$ would have been very high.

For the merger model, we take the same initial parameters as the short-t_q model and use equation (37) to calculate black hole growth, with a merging factor $f_m = 1.8$ between each pair of redshifts shown in Figure 8 ($2 \rightarrow 1$, $1 \rightarrow 0.5$, $0.5 \rightarrow 0$). This simple prescription for mergers is not fully self-consistent because the luminosity function necessarily implies some accretion growth as well. At each redshift, therefore, we calculate the mean accretion rate adopted to produce the observed QLF and shift the black hole mass function by the corresponding amount. With this additional accretion growth, M_* increases by a factor of two over each of the three redshift intervals, and by a factor of eight over the full range $z = 2-0$. The merger model's optical luminosity function is shown by long dashed lines in Figure 8. Matching the observations requires steep evolution of $p(\dot{m})$ to compensate for the large amount of black hole growth from mergers. The doubling of M_* between $z = 0.5$ and $z = 0$ leaves the merger model with a high amplitude tail of luminous systems at $z = 0$.

The goal of our ADAF model, shown by the dot-dashed lines in Figure 8, is to illustrate a case in which black holes experience substantial growth through low efficiency accretion at low redshift. This requires a high probability of having $\dot{m} < \dot{m}_{\rm crit}$, which is difficult to achieve while staying consistent with the observed QLF. In particular, we are unable to find an acceptable fit to the optical QLF by simply adjusting the parameters $\dot{m}_*$ and a of our usual double power-law $p(\dot{m})$. After some experimentation, we settled on a model with the combination of a steeper low-$\dot{m}$ slope, $a = -1.5$, and a boost to the probability of accretion rates below $\dot{m}_{\rm crit}$ by a factor of sixteen. We slightly increased M_* to $1.25 \times 10^9 M_\odot$ to improve the match to the QLF break given our changed $p(\dot{m})$, and we reduced $n_* M_*$ to 3×10^{-5} Mpc^{-3} so that there would be more overall growth of $n(M)$. With these choices, $\rho_{\rm bh}$ grows by a factor of about six from $z = 2$ to $z = 0$, and roughly two-thirds of this growth comes from objects in an ADAF mode. The optical QLF is still significantly different from that of the other models, but mostly at luminosities below the observed range. At $z = 2$, the increased number density of low luminosity objects is mainly due to the steep slope of $p(\dot{m})$, which produces many faint thin-disk systems, but at lower redshifts the ADAF mode is directly responsible for this excess of faint systems. The fact that we had to adopt such an artificial $p(\dot{m})$ to obtain a model that is even approximately consistent with the optical QLF already suggests that low-z ADAF growth of black hole masses is not important in the real universe, but it is interesting to explore the predictions of such a model nonetheless.

5.2. Black hole mass functions

We expect the black hole mass function $n(M, z)$ to be a good discriminant among our models because they involve different amounts of black hole growth, and in some cases start from different $n(M)$ at $z = 2$. The best prospects for measurements of $n(M)$ are at $z = 0$, using the observed distribution of bulge luminosities or velocity dispersions and the observed correlation of dynamical black hole masses with these properties. A number of authors have estimated the black hole mass function in this way (e.g., Salucci et al. 1999; Merritt & Ferrarese 2001; Yu & Tremaine 2002; Aller & Richstone 2002), and improving determinations of the form and scatter of the $M_{\rm bh} - \sigma$ relation (Gebhardt et al. 2000; Merritt & Ferrarese 2000) and of the distribution of bulge dispersions (Sheth et al. 2003) should yield more accurate estimates in the near future. Recent estimates of the average black hole mass density at $z = 0$ are $\rho_{\rm bh} = 2 - 3 \times 10^5 (h/0.7)^2$ M$_\odot$ Mpc^{-3} (Yu & Tremaine 2002; Aller & Richstone 2002). High redshift estimates of $n(M)$ will be difficult, since achievable angular resolution is not sufficient to measure dynamical masses of quiescent black holes. The distribution of masses of *active* systems at a given luminosity can be measured using reverberation mapping or emission line widths, and we discuss its diagnostic power in §5.4 below. It may be possible to estimate the underlying $n(M)$ by establishing a correlation between $M_{\rm bh}$ and host galaxy properties using active systems, then proceeding as at low redshift, but the observational uncertainties are likely to remain considerable.

We show the model black hole mass functions in Figure 9, but to make differences more visible we divide through by the prediction of the short-t_q model and plot the logarithm of the ratio, $\Delta \log n(M) = \log n(M) - [\log n(M)]_{{\rm short}-t_q}$. The short-$t_q$ model has little growth of the black hole mass function, with a change from $M_* = 1 \times 10^9 M_\odot$ at $z = 2$ to $M_* = 1.17 \times 10^9 M_\odot$ at $z = 0$. The long-t_q model at $z = 2$ has $n(M)$ a factor of ~ 6 lower at all masses because it has the same value of M_* and a lower value of n_*. There is more growth in the long-t_q model due to a higher $p(\dot{m})$, and the values of $n(M)$ begin to catch up to the short-t_q case. Since the growth has the effect of shifting the break in the double power-law $n(M)$ to higher masses, the change in $n(M)$ is larger at the steep, high mass end and smaller at low masses, producing the characteristic kinked shape of the lines in Figure 9. By $z = 0$ the long-t_q model has overtaken the short-t_q model at high masses, but it remains below at low masses.

The final black hole mass densities for the short-t_q and long-t_q models are $\rho_{\rm bh} = \int_0^\infty M n(M) dM = 4.02 \times 10^5 M_\odot {\rm Mpc}^{-3}$ and $1.80 \times 10^5 M_\odot {\rm Mpc}^{-3}$, respectively. The mass density *added* between $z = 2$ and $z = 0.5$ is nearly the same in the two models, $\Delta\rho_{\rm bh} \approx 0.7 \times 10^5 M_\odot {\rm Mpc}^{-3}$. This agreement is expected from the Soltan (1982) argument, which implies that $\Delta\rho_{\rm bh} = \int_{z_1}^{z_2} U(t)/(\epsilon c^2) dt$, where $U(t)$ is the bolometric emissivity of the quasar population at time t. Since mass growth in both models is dominated by thin-disk systems with the Elvis et al. (1994) SED and $\epsilon_{0.1} = 0.1$, and both models reproduce the observed optical QLF, they necessarily have similar emissivity histories and mean efficiencies and therefore similar $\Delta\rho_{\rm bh}$. The long-t_q model adds significantly more mass between $z = 0.5$ and $z = 0$, partly because it has a higher optical QLF at low redshift (Fig. 8), and partly because its $\dot{m}_*$ falls below $\dot{m}_{\rm crit}$, increasing the amount of low efficiency accretion. The short-t_q model has a final $\rho_{\rm bh}$ that is high in comparison with recent estimates, though arguably in the range of their uncertainties. The long-t_q model's $\rho_{\rm bh}$ agrees well with these estimates, coming in slightly on the low side.

The obscured model starts at $z = 2$ with the same $n(M)$ as the short-t_q model, but $n(M)$

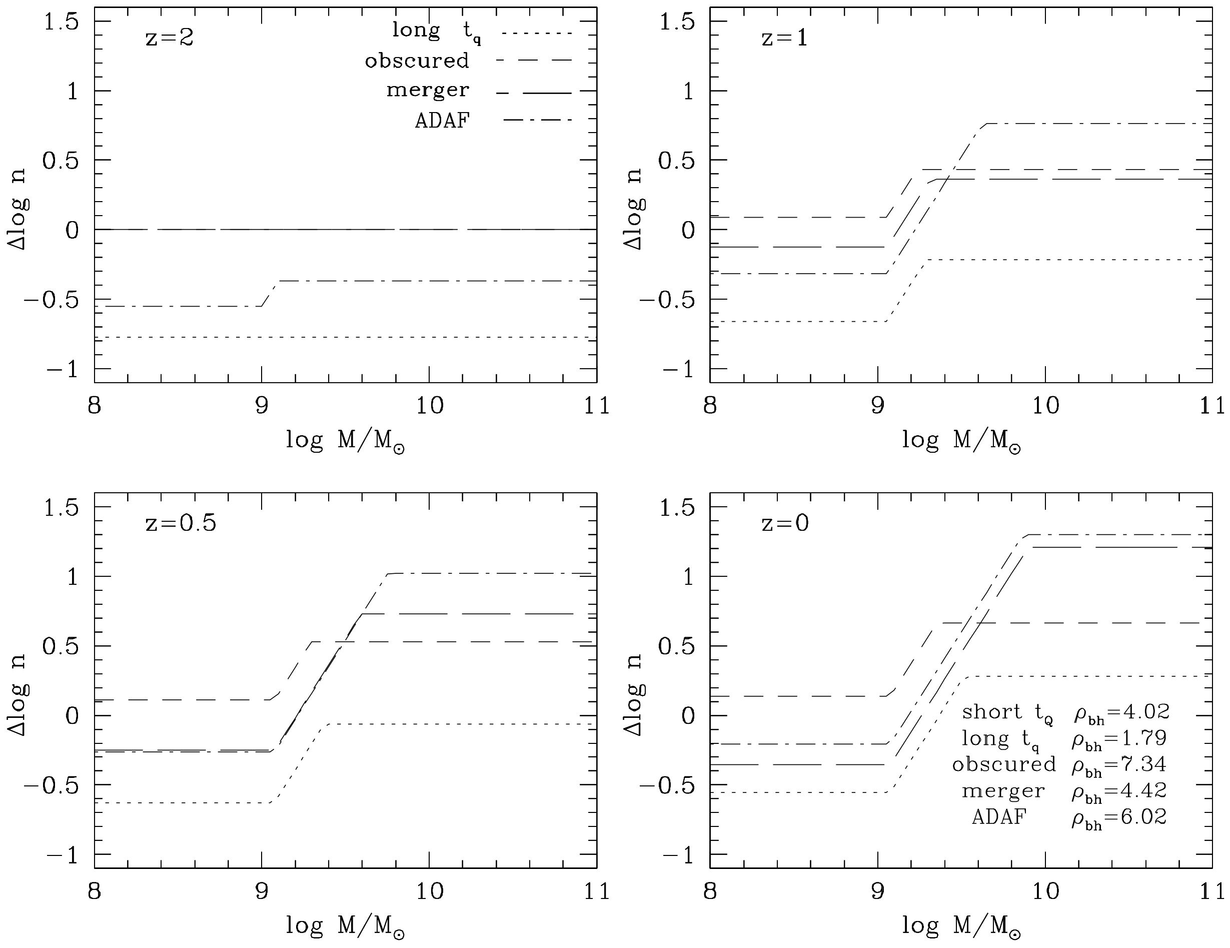

Fig. 9.— Evolution of the black hole mass function for the five models discussed in §5. Lines in each panel show $\Delta \log n \equiv \log_{10}[n(M)/n_s(M)]$, where $n_s(M)$ is the mass function of the short-t_q model at the indicated redshift and $n(M)$ is the mass function of the long-t_q, obscured, merger, or ADAF model (dotted, short-dashed, long-dashed, and dot-dashed, respectively). The $z = 0$ panel also lists the final value of the black hole mass density for each model in units of $10^5 M_\odot$ Mpc^{-3}. All models have a double power-law mass function with slopes $\alpha = -1.5$ and $\beta = -3.4$ at all redshifts, and values of $M_*(z)$ are listed in Table 3.

grows quickly because of the large amount of optically invisible accretion. The break mass M_* and mass density $\rho_{\rm bh}$ grow by a factor of 2.25 between $z = 2$ and $z = 0$, with a final $\rho_{\rm bh}$ about a factor of two larger than the short-t_q case and outside the range of recent estimates. The amount of mass added, $\Delta\rho_{\rm bh} \approx 4 \times 10^5 M_\odot {\rm Mpc}^{-3}$, is about five times the amount for the short-t_q model, which is as expected because they have similar optical QLFs while the obscured model has four optically invisible systems for each unobscured system. Since the observed optical QLF provides a fairly natural fit to the estimate $\rho_{\rm bh}$ on its own (e.g., Yu & Tremaine 2002), it is difficult to add a large amount of hidden accretion without overrunning these estimates. Higher efficiency accretion, perhaps from spinning black holes, is one option (Elvis et al. 2002). However, part of the solution is probably that our assumption of an 80% obscured fraction at all redshifts and luminosities, taken from Comastri et al. (1995) and Fabian & Iwasawa (1999), is too extreme. Recent studies show that faint X-ray sources are generally at lower redshifts than synthesis models predict, in which case a smaller fraction of obscured sources are needed to produce the hard X-ray background (Barger et al. 2002; Ueda et al. 2003).

The merger model starts with the same black hole mass function as the short-t_q case, but it evolves primarily by merging two lower mass black holes that create one higher mass black hole. By $z = 0$, high mass black holes are more numerous than in the short-t_q case by a factor of ten, and low mass black holes are less numerous by a factor of three. The *accretion* growth in this model is still dominated by thin-disk systems, and since mergers do not add mass to the black hole population, the growth of $\rho_{\rm bh}$ tracks that of the short-t_q and long-t_q models down to $z = 0.5$. Like the long-t_q model, the merger model adds more mass at $z < 0.5$ because of its high QLF and low $\dot{m}_*$.

The ADAF model starts with $M_* = 1.25 \times 10^9$ M$_\odot$ and a normalization of $n(M)$ a factor of four lower than in the short-t_q case. The high amount of accretion at low $\dot{m}$ values, in both the thin-disk and ADAF regimes, results in more black hole growth than in any of the other pure accretion models. By $z = 1$, $n(M)$ is similar to that of the short-t_q case at low masses, and it is a factor of ten larger at high masses. (Recall that all masses grow by the same factor even in the ADAF model, so this difference just reflects the slope of the mass function in the two regimes.) The final black hole mass density is 50% larger than that of the short-t_q model, and it is difficult to make parameter changes that significantly reduce this value because of the emissivity argument; the optical QLF of this model traces the observations at observed luminosities, but it has a high amplitude tail at low luminosity, and the mean efficiency is low because of the high fraction of ADAF accretion.

Figure 10 compares the $z = 0$ black hole mass functions of the five models to an estimate derived by combining the Sheth et al. (2003) estimate of the distribution of early-type galaxy velocity dispersions with the $M - \sigma$ relation found by Tremaine et al. (2002), $\log M_{\rm bh}/M_\odot = 8.13 + 4.02 \log(\sigma/200 \text{ km s}^{-1})$. Filled triangles show the case where there is no intrinsic scatter in the $M - \sigma$ relation, while filled circles and squares show results assuming a log-normal $p(M|\sigma)$ distribution with intrinsic scatter of 0.25 dex and 0.5 dex, respectively. Above $10^9 M_\odot$, the derived

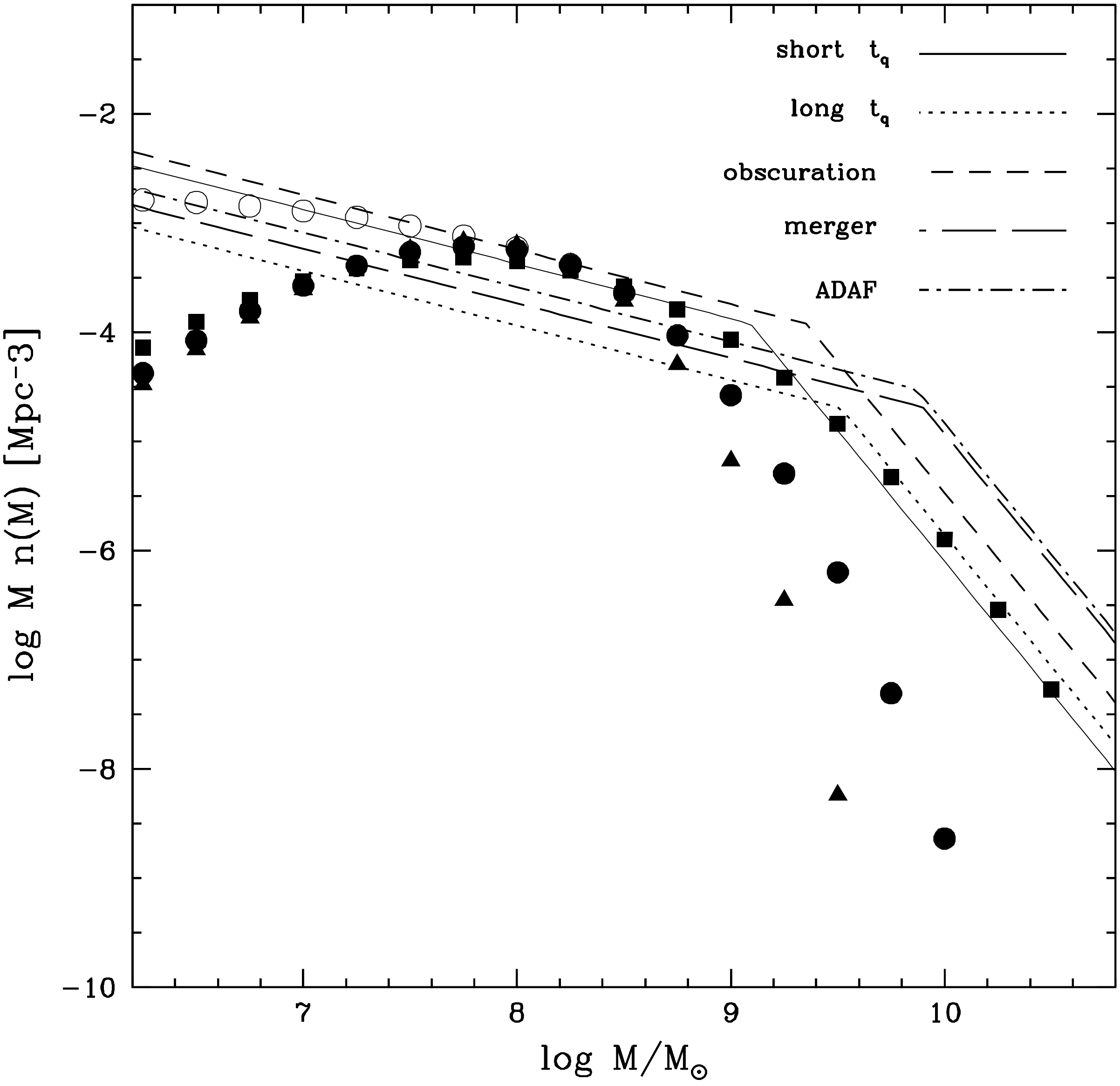

Fig. 10.— The black hole mass function at $z = 0$. Lines represent the five model mass mass functions. Solid points show the mass function derived by combining the Sheth et al. (2003) velocity dispersion distribution of early-type galaxies with Tremaine et al.'s (2002) estimate of the $M - \sigma$ relation between black hole mass and bulge velocity dispersion. Triangles, circles, and squares show results assuming no intrinsic scatter in the $M - \sigma$ relation, 0.25-dex scatter, and 0.5-dex scatter, respectively. Open circles show the effect of adding Aller & Richstone's (2002) estimated contribution from spiral galaxies to the 0.25-dex scatter (filled circle) estimate for early-type galaxies.

mass function is quite sensitive to the assumed intrinsic scatter. Tremaine et al. (2002) argue that this scatter is no larger than $\sim 0.25 - 0.3$ dex, but the data are concentrated in the range $10^7 - 10^{8.5} M_\odot$, and the scatter could increase or decrease with mass. Above $10^{9.5} M_\odot$, the derived mass function also depends on extrapolating the mean $M - \sigma$ relation beyond the range of current observations. Since Sheth et al. (2003) consider only early-type galaxies, we show with open circles the effect of adding Aller & Richstone's (2002) estimate of the spiral galaxy contribution, which becomes important below $\sim 10^{7.8} M_\odot$. If we included their S0 contribution as well (which might double count galaxies already in the Sheth et al. sample), then the mass function would be higher by 0.2 dex in this regime.

If the extrapolation of the $M - \sigma$ relation is correct and the scatter is indeed $\lesssim 0.3$ dex, then even the short-t_q model overpredicts $n(M)$ above $10^9 M_\odot$, and our other models fare worse because of their higher values of $M_*(z = 0)$. Bringing the models in line with this estimate of $n(M)$ would require either reducing our initial $M_*(z = 2)$ by 0.5-1 dex or changing our double power-law form of $n(M)$. We selected $M_*(z = 2) = 10^9 M_\odot$ because the Eddington luminosity lM_* is then close to Boyle et al.'s (2000) break luminosity (more precisely, $lM_* = 1.7 L_{\rm brk}$), making it straightforward to fit the observed $\Phi(L)$. However, given the interplay between $n(M)$ and $p(\dot{m})$, there is at least some room to reduce M_* and continue to match the observed QLF without requiring super-Eddington luminosities. If we instead adopted an exponential high-M cutoff (or a steeper high-M power-law slope), then we would need a mass-dependent $p(\dot{m})$ to reproduce the Boyle et al. (2000) data at $z = 2$, with more massive black holes having a higher probability of being active. We will explore these implications, and the tradeoff with uncertainties in the data, in future work, where we run models backwards from the $z = 0$ mass function instead of forwards from the $z = 2$ QLF.

5.3. X-ray and FIR luminosity functions

In three of our models (short-t_q, long-t_q, and merger), the bolometric emission of the quasar population is dominated by systems with a thin-disk SED. We therefore expect them to make similar predictions for the QLF in all wavebands (since they match in the B-band by construction). However, the obscured and ADAF models have large populations of systems with different SED shapes, and they may be distinguished by their X-ray or FIR luminosity functions. Observationally, the X-ray luminosity function is not as well characterized as the optical luminosity function, especially at high redshifts, but large surveys following up sources from *ROSAT*, *ASCA*, *Chandra*, and *XMM-Newton* are transforming the situation and yielding much better constraints on the evolution of the X-ray QLF (e.g., Miyaji, Hasinger, & Schmidt 2001; Cowie et al. 2003; Fiore et al. 2003; Hasinger 2003; Steffen et al. 2003; Ueda et al. 2003). In the mid/far-IR, *SCUBA* detects the brightest sources at 850μm (Priddey et al. 2003), but the revolutionary instrument should be *SIRTF*, with a much greater combination of wavelength range and sensitivity than previously available.

We calculate X-ray and FIR luminosity functions of our five models using the F_ν values in

Table 1 for systems with the various accretion modes. For X-ray QLFs we use *observed-frame* bandpasses of $0.5 - 2$ keV and $2 - 10$ keV at redshifts $z = 2$, 1, and 0.5. Results are shown in Figure 11. To enhance the visibility of model differences, we again plot the log of the ratio of each model's predictions to those of the short-t_q model. The short-t_q model's predictions at $z = 2$ are nearly identical to those shown by the solid lines in Figure 3. As expected, results for the long-t_q and merger models are very similar to short-t_q in all bands because of the dominance of thin-disk SEDs, and they could probably be brought closer still with slight adjustments of model parameters. The low-z X-ray luminosity functions of the long-t_q and merger models are slightly enhanced at low luminosities relative to B-band because their low $\dot{m}_*$ values (an indirect consequence of higher M_*) lead to more ADAF accretion.

The obscured model, which initially has five times as many accreting systems as the short-t_q model, shows nearly this full factor of five enhancement in the $2 - 10$ keV band at $z = 2$, where obscuration is almost negligible. This enhancement shrinks steadily towards lower z, especially at higher luminosities, as the observed-frame $2 - 10$ keV band becomes more affected by obscuration (see Table 1). The $0.5 - 2$ keV band shows a significant (0.4-dex) enhancement at low luminosities at $z = 2$, comprised of systems that have high bolometric luminosity and are thus able to shine detectably at this wavelength despite obscuration. However, at lower redshift the $0.5 - 2$ keV band is almost completely extinguished, and the obscured QLF is no different from that of the short-t_q model. The most dramatic feature of the obscured model is the booming FIR luminosity function, enhanced by $1 - 2$ orders of magnitude at all redshifts because the numerous obscured systems re-radiate all of their absorbed UV and soft X-ray luminosity in the FIR. This distinctive prediction of models with a large obscured population should be easily testable with *SIRTF*. The prediction holds regardless of whether obscured and unobscured systems represent two separate populations or different orientations of the same population, provided that the absorbed energy is indeed re-radiated.

As previously noted, the ADAF model has some substantial differences from the short-t_q model even in B-band. At $z = 2$, the steady rise in low luminosity systems is a consequence of the steeper slope of $p(\dot{m})$, while the bump at $\log L \leq 10^{43.5} \mathrm{erg\,sec}^{-1}$ reflects the boosted number of objects with $\dot{m} < \dot{m}_{\mathrm{crit}}$ and thus consists of objects accreting in an ADAF mode. At low redshifts, this ADAF bump moves to slightly higher luminosities. The B-band differences reappear at other wavelengths, but there are further differences in the X-ray bands that reflect the larger fraction of the ADAF SED that emerges in these bands. At high redshift, the low luminosity boost to the QLF is larger in X-ray than in optical, exceeding an order of magnitude. At low redshift, the decreasing value of $\dot{m}_*$ leads to a still higher probability of ADAF accretion, and high mass black holes in ADAF mode boost the X-ray QLF even at high luminosities. Indeed, we can infer from Figure 11 another prediction of our ADAF model: at $z \lesssim 1$, most X-ray selected AGN should be ADAF systems, at every luminosity. This does not appear to be the case in the real universe, providing a further observational argument against the importance of low efficiency accretion as a black hole growth mechanism at low redshift.

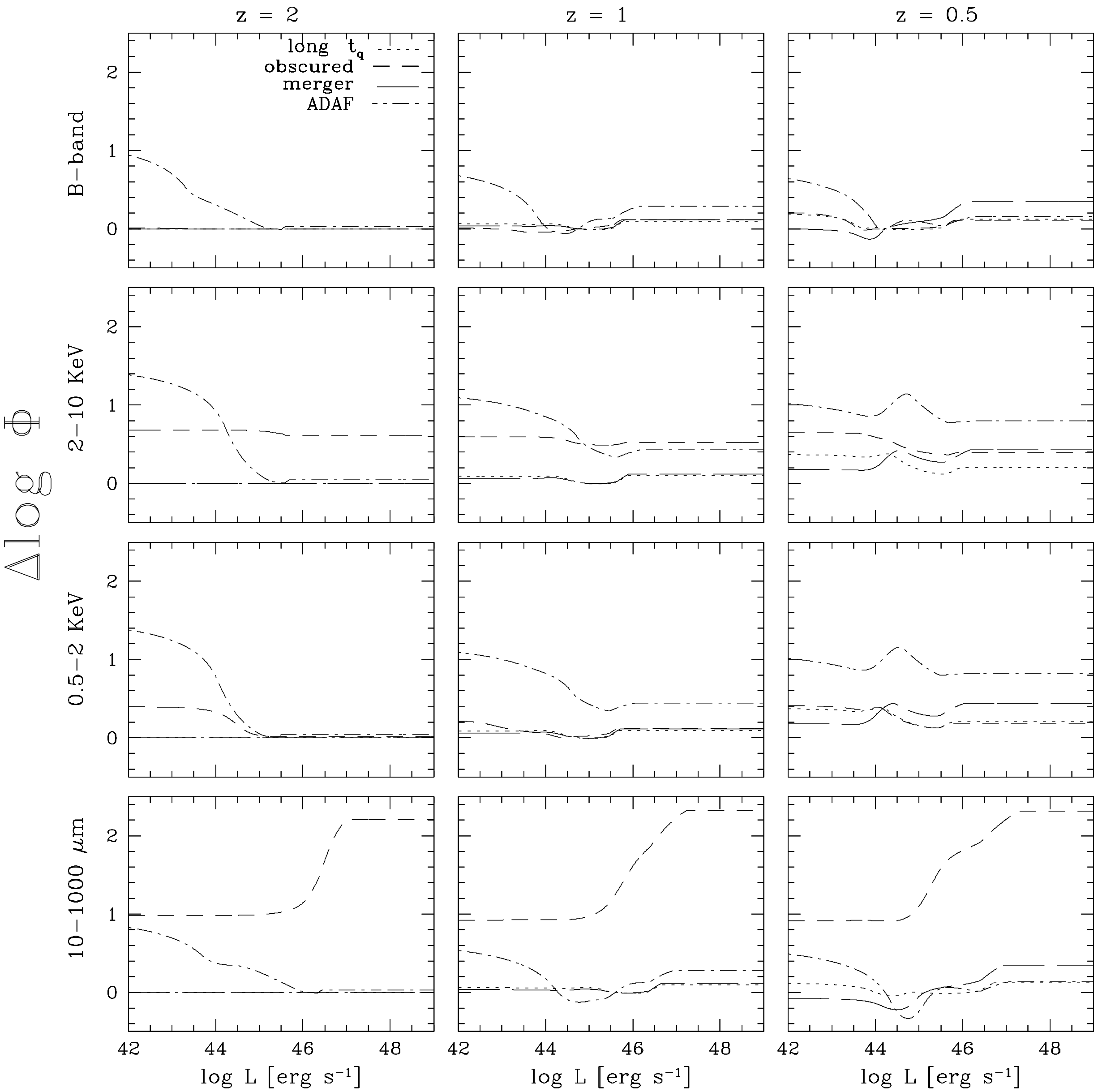

Fig. 11.— Ratios of the luminosity functions of the long-t_q, obscured, merger, and ADAF models to those of the short-t_q model at $z = 2, 1$, and 0.5 (left, middle, right). Lines in each panel show $\Delta \log \Phi \equiv \log_{10}[\Phi(L)/\Phi_s(L)]$ for B-band, 2-10 KeV, 0.5-2 KeV, and 10-1000 μm (top to bottom). The X-ray bands are observed-frame with the redshift effects on F_ν shown in Table 1 included.

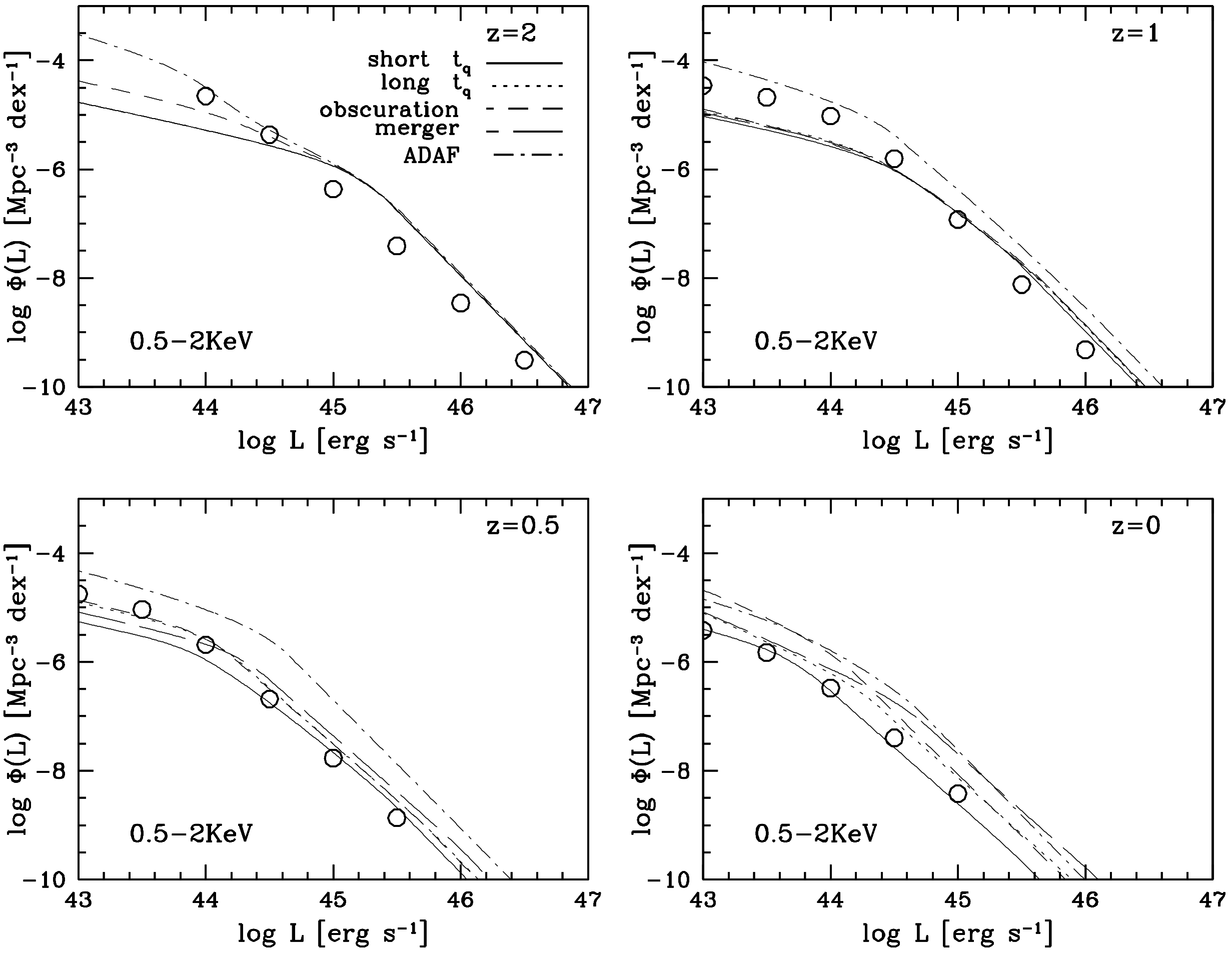

Fig. 12.— The observed frame soft X-ray luminosity function at z =2,1,0.5, and 0. Lines show the predictions of our five models, as indicated. Open circles show the evolutionary model fits to the ROSAT 0.5-2 keV QLF from Miyaji, Hasinger, & Schmidt (2001), for $\Omega_m = 1$, $h = 0.5$. Points are plotted over the range of luminosities spanned by the data at each redshift.

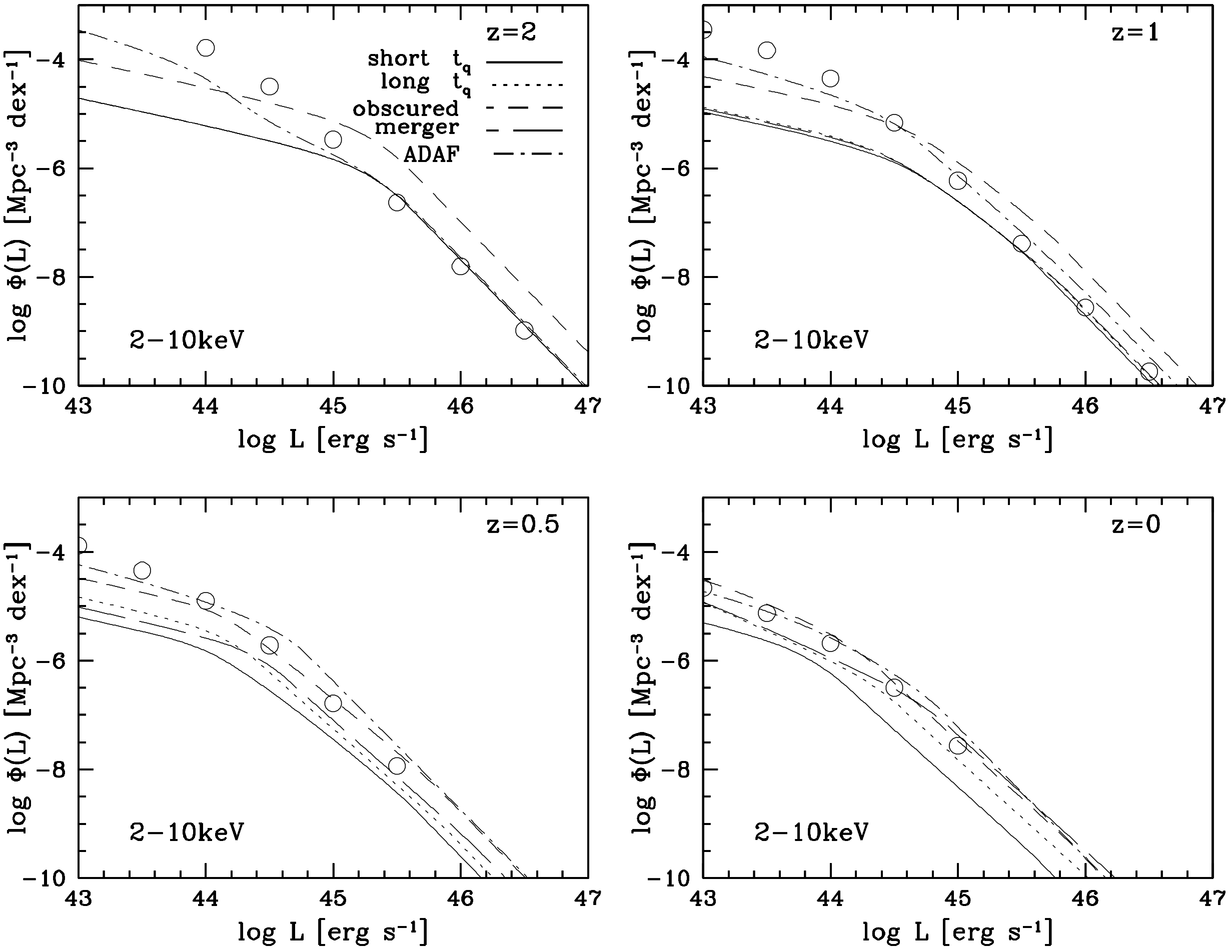

Fig. 13.— The intrinsic hard X-ray luminosity function at z =2, 1, 0.5, and 0. Lines show the predictions of our five models, as indicated. Open circles show the evolutionary model fits to the 2-10 keV QLF from Ueda et al. (2003), based on *HEAO1*, *ASCA*, and *Chandra* surveys. Ueda et al. (2003) use spectral modeling to estimate the rest-frame 2-10 keV luminosity corrected for obscuration. We therefore use the 2-10 keV F_ν values from Table 1 at each redshift, and we use the thin-disk values of F_ν for obscured systems in the obscured model, since these represent the intrinsic luminosities.

Figures 12 and 13 compare the model predictions to estimates of the soft and hard X-ray luminosity functions from Miyaji, Hasinger, & Schmidt (2001) and Ueda et al. (2003), respectively. As with the optical QLF, we plot the authors' evolutionary model fits over approximately the range covered by the observational data at each redshift. Miyaji et al.'s (2001) luminosity function, is based on observed-frame, 0.5-2 keV luminosities from the *ROSAT* All Sky Survey, with no correction for X-ray obscuration. Ueda et al. (2003), on the other hand, use spectral shape information to estimate the *intrinsic* (i.e., corrected for obscuration), *rest-frame* 2-10 keV luminosity function, from a combination of *HEAO1*, *ASCA*, and *Chandra* surveys. We compute the corresponding quantities from our models in each case.

At $z = 2$, all models fit the high luminosity end of the 2-10 keV QLF, except for the obscured model, which overpredicts by a factor of five. However, all models overpredict the 0.5-2 keV QLF by a factor ~ 3. At $z = 1$ and $z = 0.5$, models come into better agreement with the 0.5-2 keV QLF (except for the ADAF model, which remains high), but the thin-disk dominated models (short-t_q, long-t_q, merger) fall below at 2-10 keV. By $z = 0$, the M_* growth in the long-t_q and merger models has brought them back into better agreement at 2-10 keV, but they are high at 0.5-2 keV, while the short-t_q model is about right at 0.5-2 keV and well below at 2-10 keV. The obscured model is in rough agreement with both X-ray QLFs at $z \leq 1$.

In brief, the situation is confusing, and there is no single obvious change that would bring any of the models into agreement with both X-ray luminosity functions and the B-band luminosity function (Fig. 8) at all redshifts. A more sophisticated obscuration model, with soft X-ray and optical obscuration becoming more important at low redshifts and low luminosities, would certainly help, and it might explain why the faint end of the 2-10 keV QLF rises above the B-band QLF at $z = 2$. However, the overprediction of the soft X-ray QLF at $z = 2$ by models that match the B-band and 2-10 keV QLF seems difficult to understand. It is worth noting that Ueda et al.'s models, which incorporate luminosity-dependent obscuration, also tend to overpredict the soft X-ray QLF (see their figure 15), and that Miyaji et al.'s redshift bins become large at high redshift (e.g., $z = 1.6 - 2.3$ and $2.3 - 4.6$), which may make the model interpolation to a given redshift less accurate.

5.4. Masses and accretion rates of active black holes

While the underlying black hole mass function $n(M)$ may be difficult to determine at $z > 0$, the distribution of *active* black hole masses is more accessible. As discussed in §3.3, the mass distribution of active systems depends on both $n(M)$ and $p(\dot{m})$, and its variation with luminosity can be a valuable discriminant of models. Locally, the masses of active systems can be measured by combining emission line widths with sizes of the emitting regions estimated by reverberation mapping (Wandel et al. 1999). This approach can in principle be extended to high redshift, but fainter targets and longer variability timescales make it difficult. A more broadly applicable method is to combine line widths (e.g., $H\beta$, or C IV at higher redshift) with the average size-luminosity relation inferred from reverberation mapping of local objects or from photoionization modeling (e.g., Laor 1998; Gebhardt et al. 2000; McLure & Dunlop 2002; Corbett et al. 2003; Vestergaard 2004). In the last few years, these methods have yielded black hole mass estimates over an increasing range of redshift and luminosity (e.g., Woo & Urry 2002). Estimates of mass accretion rates are typically made by combining mass estimates with luminosities for an assumed efficiency, so they are not independent of the mass estimates themselves. However, quasar SEDs may also provide at least rough diagnostics of accretion rates, in the broad categorization of thin-disk vs. ADAF systems, for example, and perhaps in the finer distinction between near-Eddington systems and significantly sub-Eddington systems (Kuraszkiewicz et al. 2000; Czerny et al. 2003).

Figure 14 illustrates the range of black hole masses that contribute to different ranges of B-band luminosity at $z = 2$, 1, 0.5, and 0. For each redshift and model, a symbol marks the median mass $M_{\rm med}$ of black holes that are active in this luminosity range, and a vertical bar shows the 10%–90% range of masses at this luminosity. Small black dots show the model's value of M_*, the mass of the break in the black hole mass function. Different panels represent different luminosity ranges, which we express in terms of $L_{\rm brk}$, the break parameter in the observed B-band QLF. Bear in mind that the physical luminosity associated with $L_{\rm brk}$ decreases towards low redshift, and that $L_{\rm brk}$ drops increasingly below lM_*. Furthermore, the $z = 0$ results here and in subsequent plots should be taken with a grain of salt because we do not require our models to match the shape of the QLF at $z = 0$, only the normalization at $L_{\rm brk}$.

Figure 14 is interesting both for the features that are common to all of the models and for the features that distinguish them. The key common feature is a change in the relation between mass and luminosity from high redshift to low redshift. At $z = 2$, in all models, the sequence of luminosity is also a sequence of black hole mass: the median active mass rises from $M_{\rm med} \approx 5.6 \times 10^8 M_\odot$ in the lowest luminosity range to $M_{\rm med} = 6.8 \times 10^9 M_\odot$ in the highest luminosity range, roughly the same factor by which the luminosity itself rises. Comparing the symbols to the black dots shows that low luminosity quasars arise from sub-M_* black holes and high luminosity quasars from super-M_* black holes. The 10%–90% range of masses in a given luminosity range is only about 0.4-dex, similar to the width of the luminosity bin itself. At low redshift, on the other hand, the trend of median mass with luminosity is much weaker, and even $L > 6.25L_{\rm brk}$ systems have median masses lower than M_*. The distribution of masses for a given luminosity range is much broader than at high redshift, typically close to an order of magnitude. As we show more directly in Figure 15 below, the luminosity function at low redshift represents largely a sequence of accretion rate rather than black hole mass.

The differences between models largely trace the differences in the growth of M_*. At $z = 2$, $M_{\rm med}$ is similar at a given luminosity for all models. At lower redshifts, the short-t_q model, which has the least M_* growth, always has the lowest $M_{\rm med}$ at a given luminosity, followed by the obscured model, which has the next lowest growth of M_*. The merger and ADAF models have the most M_* growth and the highest $M_{\rm med}$ values at low redshift. The ADAF model is generally highest, even though the merger model overtakes it in M_* by $z = 0$, because the high probability of low $\dot{m}$ favors

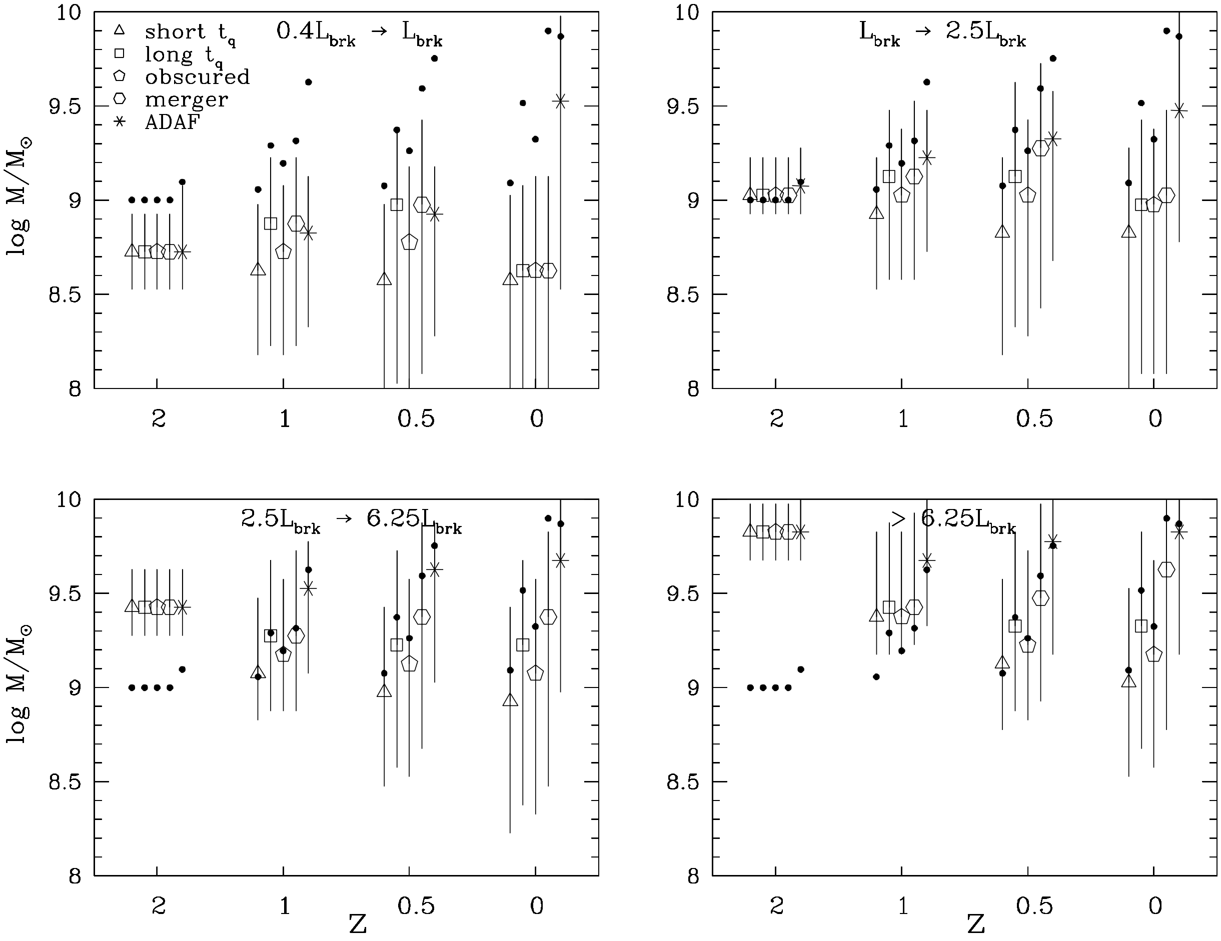

Fig. 14.— Evolution of the mass distribution of active black holes as a function of luminosity for the five illustrative models discussed in §5. Each panel corresponds to the range in B-band luminosity shown in the top center relative to the QLF break luminosity $L_{\rm brk}(z)$. The small black dots show the values of $M_*(z)$ for each model, and the open symbols show the median mass of black holes contributing to the given luminosity range for the short-t_q, long-t_q, obscured, merger, and ADAF models (triangle, square, pentagon, circle, and star). The vertical bars show the 10%-90% range of the distribution. The horizontal offset of the points from $z = 2, 1, 0.5$, and 0 is artificial and is done to distinguish the models from each other.

high mass black holes at a given luminosity. The 10%–90% spread at a given redshift is generally similar among the models.

The trends and model differences that we have shown here for B-band luminosity generally hold for luminosities defined in other bands as well. The most significant change is that the median black hole mass at fixed (in $\mathrm{erg\,sec^{-1}}$) FIR luminosity is much smaller in the obscured model than in all other models because a large fraction of the obscured SED emerges in the FIR band, enabling low mass black holes to produce high luminosity. The other significant change is that the median masses for the ADAF model are considerably lower in both X-ray bands because ADAF accretors emit a larger fraction of their bolometric energy in X-rays.

While the distributions of black hole masses and of accretion rates associated with a given luminosity provide essentially the same information, it is helpful to look directly at both distributions. Figure 15 is similar in spirit to Figure 14, but it shows the distribution of accretion rates at a given luminosity rather than the distribution of black hole masses. Symbols mark the median value of $\dot{m}$, vertical bars show the 10%–90% range of the distribution, and small black dots show each model's $\dot{m}_*$ at each redshift. A model-to-model comparison shows essentially the reverse behavior from Figure 14, with higher median black hole masses corresponding to lower median accretion rates. The key results are that the median $\dot{m}$ is close to $\dot{m}_*$ at all luminosities at $z = 2$, in all of the models, while at low redshift the median $\dot{m}$ is an increasing function of luminosity. Furthermore, even the low luminosity systems tend to have $\dot{m} > \dot{m}_*$ at low redshift.

The ADAF model is the outlier in this plot because its $p(\dot{m})$ distribution is strongly skewed to favor low accretion rates. It consistently has the lowest median accretion rate at a given luminosity and redshift, and the spread in accretion rates is large. At $z = 0.5$, the median $\dot{m}$ in the $0.4L_{\rm brk} - L_{\rm brk}$ luminosity bin is equal to the critical value $\dot{m}_{\rm crit}$ at which ADAF accretion sets in, indicating that about half of optically systems selected in this luminosity range are predicted to be ADAF accretors. A similar conclusion holds even for the highest luminosity bin at $z = 0$. The situation is more extreme for X-ray selection, where the ADAF model predicts a median $\dot{m}$ in the ADAF range in *all* luminosity ranges at $z \leq 1$. As already noted in §5.3, this prediction that a large fraction of luminous X-ray quasars are ADAF systems appears to be an empirical failure of this model.

All five of our models assume $p(\dot{m})$ independent of mass, and the transition in behavior from high redshift to low redshift is a consequence of the declining evolution of $\dot{m}_*$ that is required to match the Boyle et al. (2000) luminosity evolution in any such model. However, as shown in Figure 6, it is also possible to reproduce the Boyle et al. (2000) results with a model in which $\dot{m}_*$ is constant but $p(\dot{m}|M, z)$ is mass-dependent. Since this model is dominated by thin-disk accretors, its luminosity function is similar to that of our short-t_q model in every band. However, Figure 16 shows that its predicted distribution of active black hole masses is strikingly different at low redshift. Because the mass-dependent model reproduces the downward evolution of $L_{\rm brk}$ by preferentially reducing activity in massive systems, it predicts much lower median black hole masses at any luminosity at low redshift. Furthermore, the luminosity function remains primarily

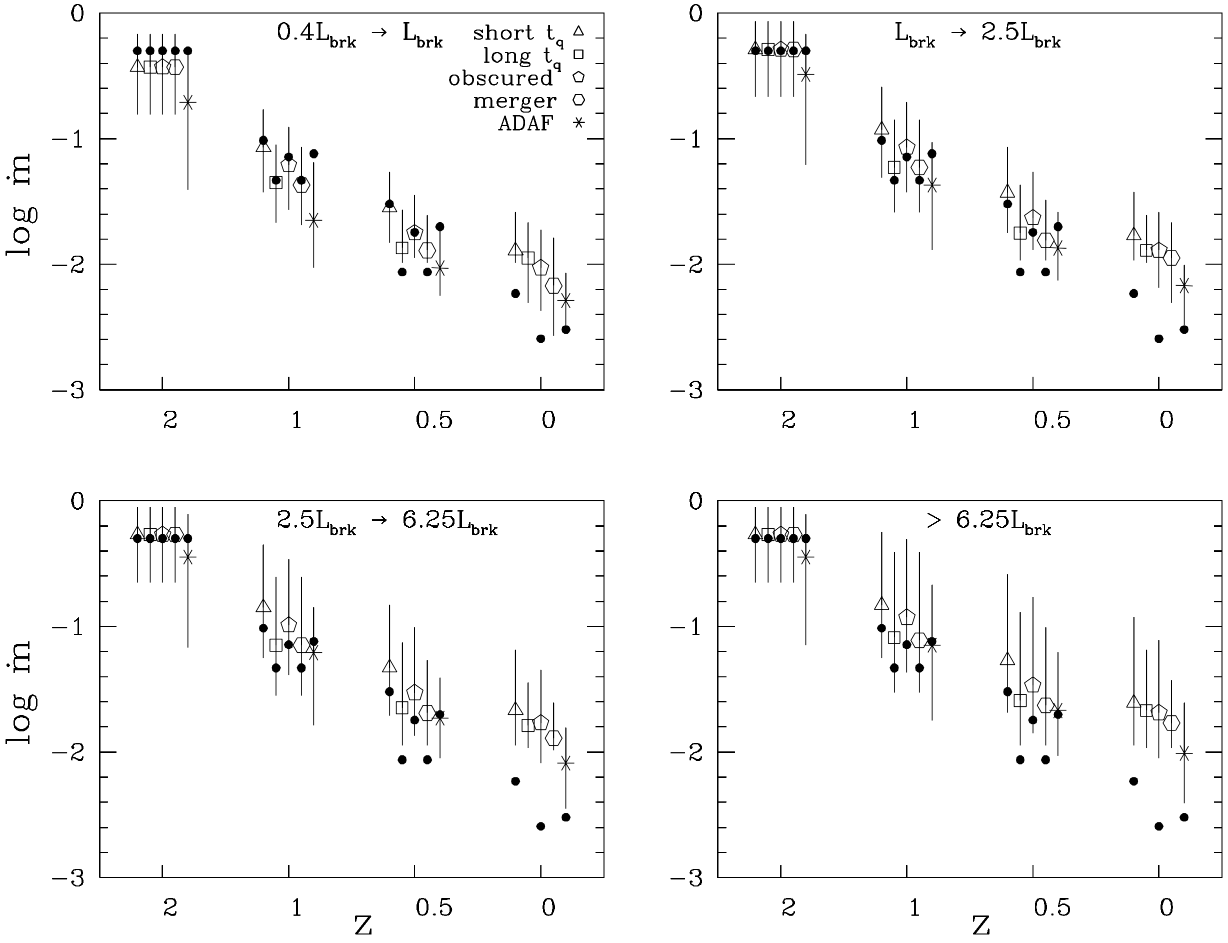

Fig. 15.— Evolution of the accretion rate distribution of active black holes as a function of luminosity for the five illustrative models discussed in §5. This figure is similar to Figure 14, but presents results in terms of accretion rates rather than black hole masses. Each panel corresponds to the range in B-band luminosity shown in the top center. The small black dots show the value of $\dot{m}_*(z)$ for each model, open symbols show the median accretion rate contributing to the given luminosity range, and the vertical bars show the 10%-90% range of $\dot{m}$. The offset of the points from $z = 2, 1, 0.5$, and 0 is artificial and is done to distinguish the models from each other. The values of $\dot{m}_*$ at $z = 0$ for the long-t_q and merger model fall well below 10^{-3} and are off the bottom of the plot.

a sequence of black hole mass even at low redshift, with the median mass rising by a factor of ten between our lowest and highest luminosity bins at $z = 0$, compared to only a factor of three for the short-t_q model. Because of the tight link between black hole mass and luminosity in this model, the spread in masses at a given luminosity remains small even at low redshift. The evolution of the distribution of active black hole masses thus provides an excellent tool for deciding whether the observed decline of $L_{\rm brk}$ reflects decreasing characteristic accretion rates or a preferential drop in activity among more massive black holes.

Recent studies using line widths to estimate black hole masses for large data samples (e.g., McLure & Dunlop 2003; Vestergaard 2004) provide the kind of data needed to test the predictions in Figure 14–16. Vestergaard's (2004) Figure 5a bins estimated black hole masses by luminosity and redshift, allowing a qualitative comparison. At $z \sim 2$, the typical estimated masses for $L \sim L_{\rm brk}$ are $\sim 10^9 M_\odot$, in agreement with our model initial conditions, and there is a clear trend of estimated black hole mass with luminosity, though perhaps less strong than the nearly linear relation predicted by our models. At $z \sim 0$, there is a broader spread in estimated masses at a given luminosity, as predicted by our standard models in which $\dot{m}_*$ declines at low z. However, the typical mass at $L \sim L_{\rm brk}$ is $\sim 10^{7.5} - 10^8 M_\odot$, which is in between the predictions of the mass-independent $p(\dot{m})$ and mass-dependent $p(\dot{m})$ models shown in Figure 16. Careful assessment and modeling of the statistical errors is needed to draw reliable conclusions from a more quantitative comparison, since the random errors in the mass estimates are large enough (a factor ~ 3) to distort the underlying mass distributions significantly.

5.5. Space densities of quasar hosts

Some of our models have a low space density of black holes and a high duty cycle — i.e., low $n(M)$ and high $p(\dot{m})$ — while others have more numerous black holes and lower duty cycles. Unfortunately, neither the luminosity function nor the distribution of active black hole masses distinguishes these cases, since both depend only on the product $n(M)p(\dot{m}|M)$ (see eqs. 8 and 21). However, if the locally observed correlations between black hole mass and bulge velocity dispersion or luminosity continue to higher redshifts, they offer a tool for diagnosing, at least approximately, the underlying space density of black holes that shine at a given luminosity. If the space density is low and the duty cycle high, then quasars should reside in rare host galaxies with luminous bulges. If the space density is high, then host galaxies should include later type and less luminous systems.

To translate this idea into a precisely defined observable, we find the median mass of black holes that produce quasars in a given luminosity range (the symbols in Figure 14), then compute the space density of black holes with this mass or greater. The corresponding observational program would require measuring the median host galaxy luminosity $L_{\rm host,med}(L_q)$ of quasars in the same luminosity range, then measuring the galaxy luminosity function at the same redshift and computing $\Phi(L > L_{\rm host,med})$, the space density of galaxies brighter than $L_{\rm host,med}(L_q)$. The predicted and observable quantities are directly comparable if the scatter between $M_{\rm bh}$ and $L_{\rm host}$ is

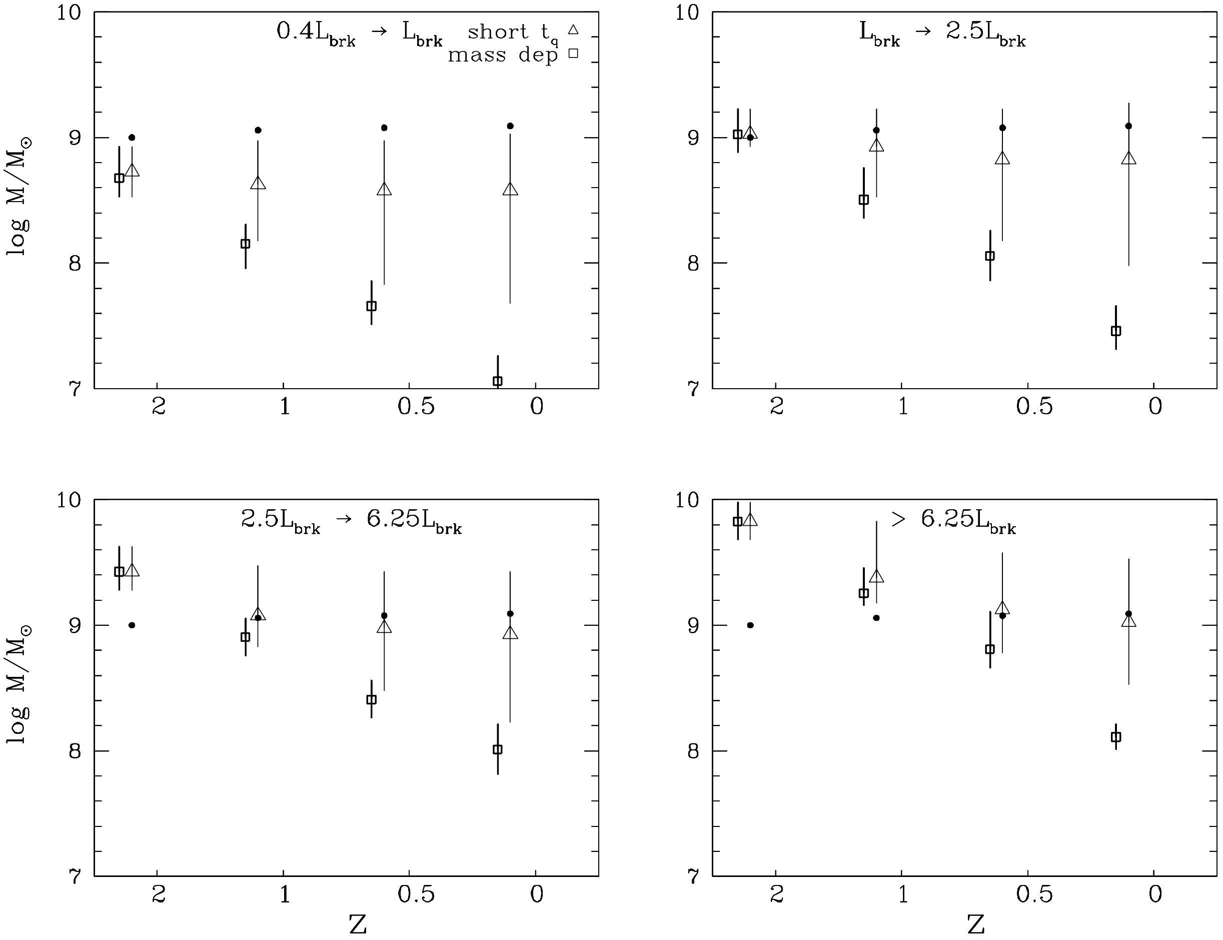

Fig. 16.— Evolution of the mass distribution of active black holes as a function of luminosity for the short-t_q model (triangles) and the mass-dependent $p(\dot{m}|M)$ model of §4.2.2 (squares). The format is similar to that of Figure 14. Small black dots show values of $M_*(z)$, which are nearly identical for both models. Open points show the median mass of active black holes in the indicated luminosity range, and vertical bars show the 10%-90% mass range.

negligible. Note that this condition does not imply negligible scatter between L_q and $L_{\rm host}$, since $\dot{m}$ variations still produce variations in quasar luminosity. Since black hole mass appears to be most directly correlated with bulge properties, one would ideally use host bulge luminosity and the bulge luminosity function rather than total luminosities. Accurate bulge-disk decomposition at high redshift may be impractical, however, especially in the presence of an active nucleus, so the next best thing is to use a red passband that is sensitive to old stellar populations. While the space density of hosts is not as informative as an actual measurement of the black hole mass function $n(M,z)$, it is less demanding observationally, since it does not require *calibration* of the $M_{\rm bh} - L_{\rm host}$ relation at high redshift, only the existence of a relation with relatively small scatter.

Figure 17 shows the model results for the luminosity ranges $0.4L_{\rm brk} < L < L_{\rm brk}$ and $L > 6.25L_{\rm brk}$ at redshifts $z = 2$, 1, 0.5, and 0. As expected, the long-t_q model starts with a host space density that is a factor of ten below that of the short-t_q model, and although it catches up at lower redshift because of the greater amount of black hole growth, a considerable gap remains. The obscured model starts with an $n(M)$ close to that of the short-t_q model, and its predictions for host space densities remain close to it at all redshifts, with the greater growth of black hole masses largely compensated by the faster decline in $\dot{m}_*$. Note, however, that this model's prediction would have been quite different if we had implemented it by boosting the black hole space density instead of the duty cycle.

The merger model is identical to the short-t_q model at $z = 2$, but its distinctive evolution of $n(M)$, with low mass black holes transforming into high mass black holes, leads to different behavior of the predicted space densities with redshift and with luminosity. In particular, the low-z depletion of the low end of $n(M)$ leads to a relatively low space density, especially at low luminosity. The ADAF model starts with a relatively low $n(M)$ (intermediate between long-t_q and short-t_q) and thus a relatively low host space density. However, its mass function overtakes that of short-t_q at high masses by $z = 1$ (see Fig. 9), and the predicted host space densities are similar for $L > 6.25L_{\rm brk}$. At low luminosities and higher redshifts, the ADAF model has a lower space density despite its large amount of black hole growth because its $p(\dot{m})$ distribution favors rarer, higher mass black holes at a given luminosity.

The model that stands out most distinctively in Figure 17, at least in terms of its redshift dependence, is the mass-dependent $p(\dot{m})$ model, shown by the filled triangles. Because this model matches the QLF by shifting activity preferentially towards low mass black holes at low redshift, the predicted host space density climbs rapidly. In this model, the hosts of moderate luminosity quasars at low redshift should include relatively late-type galaxies and low luminosity ellipticals. In effect, the properties of hosts offer an indirect way to detect the sharp decline in typical active black hole mass shown in Figure 16.

If the masses of black holes are correlated with the masses of the dark matter halos in which they reside, a natural expectation given the observed correlation with bulge velocity dispersion, then quasar clustering provides another observational tool for inferring their space density. A low black

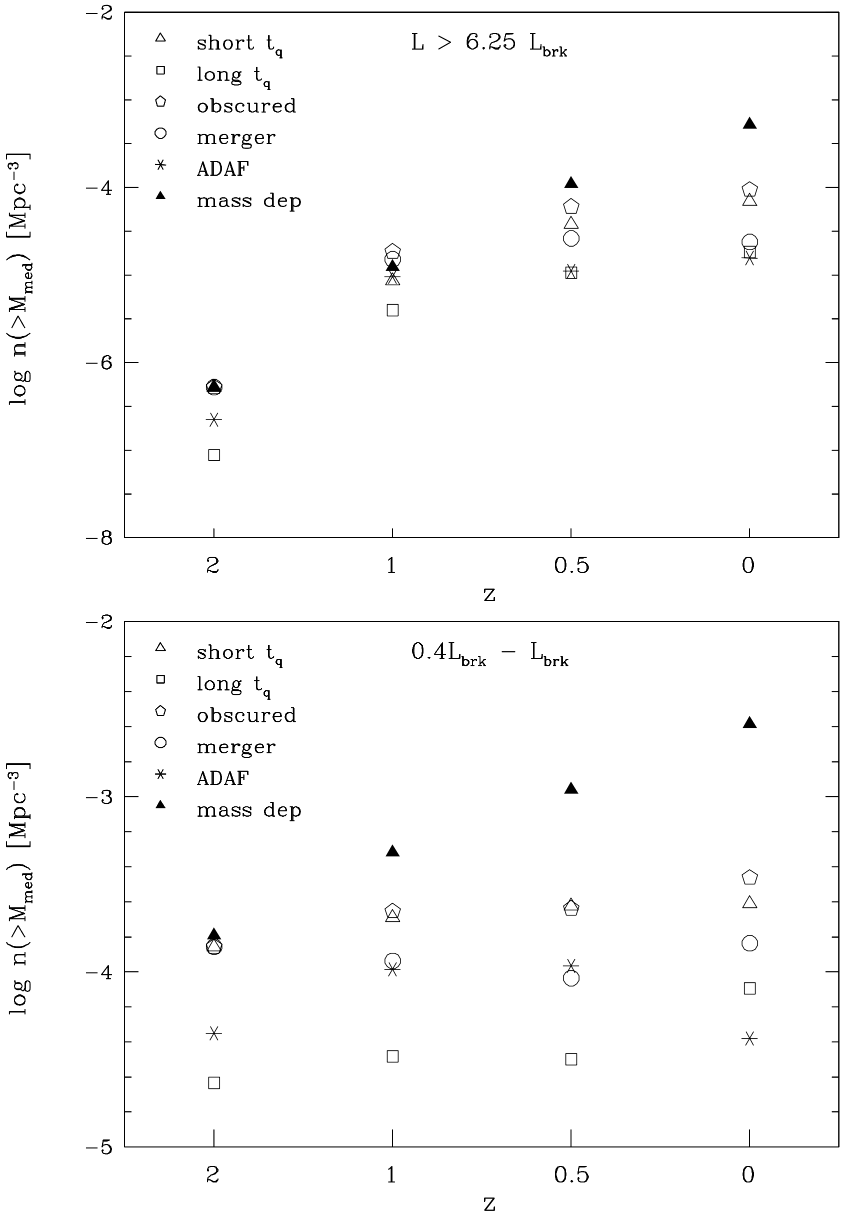

Fig. 17.— The comoving space density of black holes with mass above the median mass of active systems with B-band luminosity $0.4L_{\rm brk} < L < L_{\rm brk}$ (top) or $L > 6.25L_{\rm brk}$ (bottom). Open symbols show results for the five models discussed in §5, as indicated, and filled triangles show results for the mass-dependent $p(\dot{m})$ model illustrated in Figure 6. These predictions can be tested by studies of quasar hosts or quasar clustering. Note the different y-axis ranges of the top and bottom plots.

hole space density implies that the host halos are rare, massive systems that tend to be strongly clustered (Kaiser 1984; Mo & White 1996), while a high space density implies more common, less strongly clustered hosts. This is the idea behind the proposals of Haiman & Hui (2001), Martini & Weinberg (2001), and Kauffmann & Haehnelt (2002) to constrain quasar lifetimes (hence duty cycles, hence black hole space densities) using the quasar correlation function or the quasar-galaxy cross-correlation function (see also Haehnelt et al. 1998). In Figure 17, models with low points predict strong quasar clustering and models with high points predict weaker clustering. To a first approximation, one could calculate the expected bias (relative to mass clustering) for quasars in the two luminosity ranges by reading off the space density from the figure, finding the mass threshold for halos that have this space density given an assumed cosmological model, and calculating the bias of halos above this mass threshold using the methods of Mo & White (1996; see Martini & Weinberg [2001] for a more detailed description of this approach). One could also carry out a more thorough calculation, weighting the contribution to the bias by the fraction of black holes of a given mass contributing to the luminosity range, but we suspect that the results would not be very different.

Recent results from the 2dF Quasar Redshift Survey favor a relatively low comoving correlation length, $s_0 \approx 5.8h^{-1}$ Mpc in redshift space, with no clear evidence for dependence on redshift or luminosity (Croom et al. 2003). Using the Martini & Weinberg (2001) model, which assumes simple on-off quasar activity with a monotonic relation between quasar luminosity and host halo mass, the implied lifetime is short, $t \sim 10^6$ years. The current data set is not quite large enough to allow a precise clustering measurement for a volume-limited subset of quasars at $z \sim 2-3$, which is what one would ideally like to use for the lifetime analysis. However, if future results continue to show a low correlation length at high redshift, and no significant dependence on luminosity, then they may indicate that there is substantial scatter in the relation between quasar luminosity and host halo mass. This scatter could in turn indicate that the correlation between black hole mass and halo mass at high redshift is much weaker than the measured correlation between black hole mass and bulge mass at low redshift, or else that the luminous quasars have a wide range of $L/L_{\rm Edd}$ even at high-z. Clustering analyses of the full 2dF quasar survey and of the SDSS quasar survey should yield interesting insights on these questions over the next few years.

6. Summary

In the framework developed here, the central actor in black hole and quasar evolution is the accretion rate distribution $p(\dot{m}|M,z)$, the probability that a black hole of mass M accretes at a rate $\dot{m}$ (in Eddington units) at redshift z. Given a model for the accretion efficiency as a function of $\dot{m}$, which can be inferred from observations and theoretical considerations that are largely independent of QLF evolution *per se*, the combination of $p(\dot{m}|M)$ and the black hole mass function $n(M)$ determines the bolometric luminosity function via equation (8). Furthermore, in the absence of mergers, $p(\dot{m}|M,z)$ determines the evolution of $n(M,z)$ given a "boundary value" of $n(M)$ at some redshift. Mergers can complicate the picture by changing $n(M)$ independently of $p(\dot{m}|M,z)$. We have generally made the plausible but not incontrovertible assumption that black holes accreting in the range $0.01 < \dot{m} < 1$ have "thin-disk" efficiencies $\epsilon_{0.1} \approx 1$ and that efficiencies decrease at higher (super-Eddington) and lower (ADAF) accretion rates (eq. 7). We have derived many of our results under the mathematically simplifying assumption that $p(\dot{m})$ is independent of mass. While this assumption is unlikely to hold to high accuracy, it may be a reasonable approximation at redshifts near the peak of quasar activity, since black holes that grow faster than their peers tend to reduce their accretion rates in Eddington units and *vice versa.* Most of our specific examples assume that $n(M)$ and $p(\dot{m})$ are double power-laws with breaks at M_* and $\dot{m}_*$, respectively, with $p(\dot{m})$ truncated at $\dot{m}_{\rm min} = 10^{-4}$ and $\dot{m}_{\rm max} = 10$.

6.1. Basic Results

Our framework yields a number of mathematical results that give insight into the relations among the black hole mass function, the accretion rate distribution, and the QLF. When $p(\dot{m})$ is independent of mass, the convolution integral (8) for $\Phi(L)$ can be understood as follows: for each range $\dot{m} \rightarrow \dot{m} + d\dot{m}$, the mass function $n(M)$ is mapped to a luminosity $L = \epsilon_{0.1}\dot{m}lM$, multiplied by $p(\dot{m})d\dot{m}$, and added to a running total. If $n(M)$ is a double power-law and the range of $\dot{m}$ is bounded, then the asymptotic slopes of $\Phi(L)$ must equal the low and high mass slopes of $n(M)$, since high luminosity objects must come from black holes with $M > M_*$ and low luminosity objects must come from black holes with $M < M_*$. In the intermediate luminosity regime, where black holes above and below M_* can both contribute, the QLF turns over in a way that depends on the slopes of $n(M)$ and the shape of $p(\dot{m})$. For our usual double power-law $p(\dot{m})$, the QLF break occurs at a luminosity $L_{\rm brk} \sim \dot{m}_* l M_*$. If $\dot{m}_*$ is close to one, then the break luminosity corresponds roughly to the Eddington luminosity of M_* objects, and the slope above the break corresponds to the high mass slope of $n(M)$. However, if $\dot{m}_*$ is low, then the turnover may be associated largely with the change in slope of $p(\dot{m})$, and the asymptotic regime where the high-L slope of $\Phi(L)$ matches the high-M slope of $n(M)$ may only be reached beyond the observed range of luminosities. For most of the models that we have presented here, which are designed to match the Boyle et al. (2000) optical luminosity function, the first case applies at high redshift and the second at low redshift.

A key feature of our model of accretion physics is that objects do not radiate at super-Eddington luminosities — instead, we assume that the efficiency is $\epsilon_{0.1} = \dot{m}^{-1}$ for $\dot{m} > 1$, so that super-Eddington accretors radiate at $L = L_{\rm Edd}$. Although we have generally adopted $\dot{m}_* = 0.5 - 1$ for fitting data at $z \geq 2$, our results would not be very different if we took $\dot{m}_* > 1$, or if we changed the maximum accretion rate or the high-$\dot{m}$ slope, because the change in efficiency would cut off the luminosity distribution at $L_{\rm Edd}$ anyway. Our results would also not be very different if we assumed that black holes fed at a super-Eddington rate by their host galaxy regulate their accretion by outflows or convection (Blandford & Begelman 1999; Quataert & Gruzinov 2000) so that they gain mass at $\dot{m} \approx 1$ and radiate at $L \approx L_{\rm Edd}$. If such flows drove gas out into the galaxy halo, they would effectively truncate $p(\dot{m})$ at $\dot{m} = 1$, while if they returned gas to a reservoir from which it would eventually be accreted, they would transform the $\dot{m} > 1$ tail of $p(\dot{m})$ into a spike at $\dot{m} = 1$. With our standard efficiency assumptions, the latter scenario would increase the average radiative efficiency of high-$\dot{m}$ accretion, by moving it from the super-Eddington regime to the thin-disk regime, thus yielding more luminosity for a given amount of black hole growth. Changing the accretion physics to allow $\epsilon_{0.1} \approx 1$ with $\dot{m} > 1$, and thus to allow substantially super-Eddington luminosities when the accretion rate is high (Begelman 2002), would have a more drastic effect on our results. In this case, only a truncation of $p(\dot{m})$ would cut off the luminosity distribution at a given black hole mass, and the predicted luminosity function would therefore be sensitive to the values of $\dot{m}_*$ and $\dot{m}_{\max}$ and to the shape of $p(\dot{m})$ at $\dot{m} > 1$.

The fraction of black holes of mass M that are active at a luminosity L is proportional to $n(M)p\left(\dot{m} = \frac{L}{\epsilon_{0.1}\dot{m}lM}\right)$, the underlying black hole space density times the probability of having the accretion rate required to shine at L. The same product appears in the integrand for the luminosity function itself (eq. 8), but by constraining the integrand rather than the integral, measurements of active black hole masses can discriminate among models that produce similar $\Phi(L)$ with different $p(\dot{m})$ and $n(M)$. In terms of our double power-law models, the active mass distribution is particularly useful for distinguishing cases with high M_* and low $\dot{m}_*$ from models with low M_* and high $\dot{m}_*$.

A general mass-dependent $p(\dot{m})$ can be written in the form $p(\dot{m}|M) = p_0(\dot{m})D(M|\dot{m})$, with $p_0(\dot{m}) \equiv p(\dot{m}|M_0)$ for some fiducial mass M_0 and $D(M_0|\dot{m}) \equiv 1$. With such a mass dependence, we can still understand the convolution integral for $\Phi(L)$ by considering the contribution from each range $\dot{m} \rightarrow \dot{m} + d\dot{m}$, but where the sum before involved only horizontally and vertically shifted versions of the mass function $n(M)$, now the mass function can be tilted or distorted by $D(M|\dot{m})$ before being added to the running total. Mass-dependence of $p(\dot{m})$ thus breaks the tight link between $n(M)$ and $\Phi(L)$ and adds freedom to models of the QLF. Specifically, for a model with a given $n(M)$ and a mass-independent $p(\dot{m})$, there is a family of models with different $n(M)$ and mass-dependent $p(\dot{m})$ that yield identical predictions for $\Phi(L)$ and the distribution of *active* black hole masses. However, this degeneracy applies only at a single redshift; the evolution of models with mass-dependent $p(\dot{m})$ is different because the shape of $n(M)$ changes with time. Furthermore, while the active black hole mass distributions are the same, the underlying black hole mass functions are different, and a measurement of the full $n(M)$ at $z = 0$ may be sufficient to diagnose the mass-dependence of $p(\dot{m})$ at higher redshifts. If two models have similar underlying $n(M)$, then the masses of active black holes or space densities of their host galaxies are powerful diagnostics for mass dependence of $p(\dot{m})$, as illustrated in Figures 16 and 17.

For evolutionary calculations, a crucial simplification (eq. 10) is that the accretion driven growth of $n(M)$ depends only on the mean accretion rate $\langle\dot{m}(M,z)\rangle \equiv \int_0^\infty \dot{m}p(\dot{m}|M,z)d\dot{m}$, not on the full form of the distribution function. Between times t_1 and t_2, a black hole with accretion rate $\dot{m}(t)$ grows in mass by a factor $\exp(t_{\rm acc}/t_s)$, where $t_{\rm acc} = \int_{t_1}^{t_2} \dot{m}dt$ is the accretion weighted lifetime (eq. 23) and t_s is the Salpeter time (eq. 5). In any particular mass bin, some of the black holes grow faster than average and some grow slower, but $n(M)$ evolves as if all of them grew at the average rate for that bin. If $\langle \dot{m} \rangle$ is independent of mass, then the growth factor is the same for all mass bins, and the mass function simply shifts in a self-similar fashion, preserving its shape (eq. 11). One might expect this self-similar behavior to emerge, approximately, as a "fixed point" solution during the epoch of rapid black hole growth. If $\langle \dot{m} \rangle$ depends on mass, then the shape of $n(M)$ changes with time, becoming steeper if low mass black holes grow more rapidly and shallower if high mass black holes grow more rapidly.

Our results also imply a number of critical values for the logarithmic slopes of $p(\dot{m})$ and $n(M)$, where the character of the solutions changes. We summarize these critical slopes in Appendix B. The most important result of this sort is that growth by accretion drives $n(M)$ at fixed mass up if the mass function is steeper than M^{-1} and down if it is shallower, and that the corresponding critical slope for merger driven growth is -2 rather than -1. Generically, one expects mergers to increase $n(M)$ at high masses and decrease it at low masses.

6.2. Implications of the Observed Luminosity Evolution

We have kept the comparison to observations in this paper rather loose, saving detailed tests of models against multi-wavelength luminosity functions and estimates of black hole mass distributions for future work. However, we are able to draw some interesting general conclusions from our efforts to reproduce the observed evolution of the optical luminosity function.

At $z \leq 2$, we have focused on the Boyle et al. (2000) evolution results, and particularly on their finding that the QLF has a break that shifts to lower luminosities at lower redshifts. In any model where $p(\dot{m})$ is independent of mass, reproducing this result requires a shift towards lower typical accretion rates over time (Fig. 5). Decreasing the duty cycle uniformly by reducing the amplitude of $p(\dot{m})$ can lower the normalization of $\Phi(L)$, but on its own it cannot shift the break to lower luminosities, since black hole masses, and thus the location of the break in $n(M)$, can only increase with time. Lowering the break luminosity requires decreasing the probability of high accretion rates *relative* to low accretion rates. In the context of our double power-law models, this change is achieved by reducing $\dot{m}_*$, from slightly below unity at $z \sim 2$ to far below unity at $z \sim 0 - 0.5$. Physically, such a change could arise because of decreasing gas supplies and increasing dynamical times in galaxies at lower redshifts (Kauffmann & Haehnelt 2000).

A consequence of this reduction in $\dot{m}_*$ is a change in the nature of the QLF between high and low redshift. At high redshift, the sequence of quasar luminosities is primarily a sequence of black hole masses. Because $\dot{m}_*$ is close to unity and $p(\dot{m})$ is shallow below the break, most high luminosity quasars are produced by the more numerous low mass black holes accreting at $\dot{m} \approx 1$ rather than extremely rare high mass black holes with low accretion rates. The typical mass at a given luminosity is $M \approx L/l$, and a factor of ten increase in luminosity roughly corresponds to a factor of ten increase in black hole mass. However, once $\dot{m}_*$ falls well below unity, there is a large range $\dot{m}_* < \dot{m} < 1$ over which $p(\dot{m})$ is steep ($\propto \dot{m}^{-3}$ for most of our models), thus increasing the probability that a given luminosity is generated by a high mass black hole accreting at a low $\dot{m}$ close to $\dot{m}_*$. The typical active black hole mass still increases with luminosity, but at low redshift a factor of ten increase in L corresponds to only a factor ~ 3 increase in median black hole mass, and the range in black hole mass at fixed luminosity is much larger than at high redshift. At low redshift, the sequence of quasar luminosities remains partly a sequence of black hole mass, but it is also in large part a sequence of $\dot{m}$ (Figs. 14 and 15).

This change in character arises only because we reduce the radiative efficiency in the super-Eddington regime, forcing systems with $\dot{m} > 1$ to radiate at the Eddington luminosity. Without this transition in the accretion physics, there would be no preferred scale to change the relative importance of mass and $\dot{m}$ between high and low redshift, since we assume the same double power law *form* of $p(\dot{m})$ at all redshifts and it is only the value of $\dot{m}_*$ relative to unity that changes. If the efficiency does not decrease in the super-Eddington regime, then the shape of the luminosity function depends in detail on the form and cutoff of $p(\dot{m})$ at $\dot{m} > 1$. However, the limited data on masses of high redshift black holes generally shows systems with $L/L_{\rm Edd} \approx 1$ but not substantially larger (McLure & Dunlop 2003; Vestergaard 2004). This result suggests that there is indeed a break in the radiative efficiency at $\dot{m} \approx 1$, or else that black hole accretion self-regulates to enforce $\dot{m} \lesssim 1$, since otherwise it would require $p(\dot{m})$ to cut off coincidentally at $\dot{m} > 1$ and be relatively high at $\dot{m}$ just below one.

If $p(\dot{m})$ is mass-dependent, then there is a fundamentally different alternative for explaining the shift of the QLF break to lower luminosities: instead of a decline in characteristic accretion rates, the mass-dependence can itself evolve so that activity is preferentially suppressed in high mass black holes at low redshift (Fig. 6). Physically, such behavior could arise because high mass black holes reside in early type galaxies with large bulges, which tend to exhaust their gas supplies earlier. In this scenario, quasars at a given position on the luminosity function (relative to $L_{\rm brk}$) are always associated with the same distribution of accretion rates, but the associated black hole mass declines with redshift as the high mass black holes turn off. In comparison with the mass-independent $p(\dot{m})$ models, the median black hole mass at fixed luminosity is lower, and the range of black hole masses is smaller, with near-Eddington accretors making a large contribution to the high end of the luminosity function at all redshifts (see Figure 16). The low black hole mass implies a high space density of hosts, so this model predicts that a larger fraction of low redshift quasars reside in late type or low luminosity galaxies.

Based on anecdotal evidence, it is hard to say whether a decline in characteristic accretion rates or a decline in the activity of high mass black holes is more important in producing the observed decline of $L_{\rm brk}$ at $z < 2$. The first scenario's prediction of a wider range of $L/L_{\rm Edd}$ at lower redshifts seems in qualitative agreement with studies of black hole masses (Woo & Urry 2002; Vestergaard 2004). The second scenario seems in qualitative agreement with the relative quiescence of the black holes in most massive ellipticals (such as M87), though Kauffmann et al. (2003) find that the most luminous AGN in the local universe do reside in massive, early type systems. The two scenarios make quantitatively different predictions, and careful comparison to studies of active black hole masses and host galaxy properties should be able to show whether one mechanism dominates or both are comparably important. The results of Vestergaard (2004) suggest that active black hole masses at low redshift are intermediate between the values predicted by the two models shown in Figure 16.

At high redshifts, constraints on the form of the QLF are weaker; in particular, the SDSS measurements of Fan et al. (2001) probe only the high luminosity end of $\Phi(L)$. For matching the observed evolution over the range $z \sim 5$ to $z \sim 2$, we find one acceptable solution in which $p(\dot{m})$ is roughly constant in Eddington units and the growth of the black holes themselves drives the growth in amplitude of the QLF (Fig. 7). With our adopted parameters, the black holes grow by a factor ~ 10 between $z = 5$ and $z = 2$, and the space density at $z = 2$ is $n_* M_* = 2.012 \times 10^{-5}$ Mpc^{-3} at $M_* = 10^9 M_\odot$. Significantly lower normalizations of $n(M)$ at $z = 2$ are not allowed because they would require a negative black hole mass density at $z = 5$. Higher normalizations are allowed, in which case the quasar duty cycle is shorter, the rate of black hole growth is smaller, and the growth of the QLF is driven by a steady increase in $p(\dot{m})$. While a solution with a high black hole space density can give an acceptable match to the QLF, it seems physically unattractive because it requires that most of the black hole mass density was already in place at $z = 5$, before the main epoch of quasar activity. In this latter scenario, the luminous phases of quasars represent the addition of a small amount of mass to already formed black holes. If we assume instead that the observed optical QLF *does* trace the growth of black holes, and that the former model is therefore more realistic, then our analysis predicts a black hole space density $Mn(M) \sim 2 - 3 \times 10^{-5}$ Mpc^{-3} at $M = 10^9 M_\odot$ at $z = 2$, and continuation to $z = 0$ implies a similar space density at $M \sim 2 \times 10^9 M_\odot$. However, we have not investigated the sensitivity of this prediction to our specific choices of parameters, such as the double power-law form and adopted slopes of $p(\dot{m})$ and $n(M)$. Black hole mergers and obscured accretion could also alter the prediction significantly, especially at low redshift.

6.3. Distinguishing Scenarios

Many of the qualitative results mentioned above could have been anticipated without detailed calculations. Our framework, however, allows one to compute quantitative predictions of concrete models that illustrate distinct ideas about the nature of black hole and quasar evolution. We did this in §5 for the low redshift ($z \leq 2$) regime, adopting as our baseline a model with quasar emission and black hole growth dominated by unobscured thin-disk accretion and a normalization of $n(M, z = 2)$ implying a short quasar lifetime, and, consequently, little growth of black hole masses from $z = 2$ to $z = 0$. We compared this model to four variants: one with a lower $n(M, z = 2)$ and correspondingly longer quasar lifetime, one with a 4:1 ratio of obscured to unobscured systems, one with a large amount of merger driven growth of black hole masses, and one with a boosted probability of low-$\dot{m}$ accretion leading to substantial ADAF growth of black holes. For each scenario, we are able to find parameters that acceptably reproduce the Boyle et al. (2000) optical QLF at $z = 2$, 1, and 0.5.

Relative to the short-t_q model, the long-t_q model starts at $z = 2$ with a factor ~ 6 lower $n(M)$ at every mass. Because of the larger amount of accretion per black hole, however, the long-t_q $n(M)$ overtakes the short-t_q $n(M)$ at high masses by $z = 0$, while remaining below it at low masses. The two models have similar QLFs at every wavelength, since they match in the optical by construction and are dominated by systems with the Elvis et al. (1994) SED. However, the growth of M_* in the long-t_q model requires a more rapid decline of $\dot{m}_*$ to compensate, so at lower redshifts it predicts lower median $\dot{m}$ and higher median black hole mass at a given luminosity. The two models can thus be distinguished observationally by the $z = 0$ black hole mass function, by the mass distributions of active black holes at $z \sim 0.5 - 1$, and by the space densities of host systems, which are lower in the long-t_q model at every luminosity and redshift.

In the obscured model, the large amount of obscured accretion produces more black hole growth than in the short-t_q model, leading to a higher $n(M)$ and $\rho_{\rm bh}$ at low redshift. By $z = 0$, $n(M)$ is higher by a factor ~ 6 at high masses. However, the median black hole mass and host space density at a given optical luminosity are only slightly higher. The clearest distinguishing feature of this model is the relative amplitude of luminosity functions in different wavelength bands, a consequence of the different SED shapes of obscured and unobscured accretors. At $z = 2$, the 2-10 keV luminosity function is elevated by nearly a factor of five, the ratio of all systems to unobscured systems. This boost decreases towards low redshift because of the increasing importance of obscuration in the (observed-frame) 2-10 keV band. The 0.5-2 keV band is heavily obscured at $z < 2$, so the soft X-ray luminosity function of the obscured model is similar to that of the short-t_q and long-t_q models, except at $z = 2$ where it has a higher amplitude at low luminosity. The strongest departure of all is in the FIR, where the re-radiated emission of obscured accretors boosts $\Phi(L)$ by factors of ten (low luminosity) to one hundred (high luminosity), relative to the short-t_q and long-t_q models. This crucial prediction should soon be testable by *SIRTF* and by other sub-mm and mm-wavelength observations.

In the merger model, low redshift mergers strongly distort the initial black hole mass function, depleting it at low masses and boosting it at high masses, with a factor of 16 increase in the high mass end at $z = 0$ relative to the short-t_q model. The quasar population of this model is still dominated by systems with a thin-disk SED, and since it matches the observed optical luminosity function by construction, its predictions at other wavelengths are close to those of the short-t_q and long-t_q models. However, the high space density of high mass black holes leads to a high median mass of active black holes at fixed luminosity, similar to that of long-t_q at low luminosities and higher still at high luminosities. Merger driven distortions of the mass function also lead to distinctive redshift and luminosity dependence of the host space density.

Finding parameters that yield significant ADAF growth and an acceptable match to the optical luminosity function proves quite difficult, requiring an artificial boost to the probability of accretion rates below $\dot{m}_{\rm crit}$. With our adopted parameters, the ADAF model predicts a large amount of black hole growth between $z = 2$ and $z = 0$, and thus a high $n(M)$ and $\rho_{\rm bh}$ at $z = 0$. With the combination of high M_* and high probability of low $\dot{m}$, the ADAF model predicts the largest median black hole masses and lowest median accretion rates at fixed luminosity, with the median accretion rate approaching $\dot{m}_{\rm crit}$ even for high luminosity AGN at $z \leq 0.5$. The high X-ray fraction of the ADAF SED boosts the soft and hard X-ray luminosity functions relative to the optical, especially at low redshift, and the model predicts that a majority of X-ray selected systems at $z \sim 0.5$ should be ADAF accretors, even at high luminosities. This prediction appears observationally untenable, and the model requires rather implausible parameter choices in the first place, so our results suggest that ADAFs are unlikely to make an important contribution to black hole growth in the real universe, even at $z < 2$ (Haehnelt et al. [1998] reach a similar conclusion). Low radiative efficiency at low $\dot{m}$ may nonetheless help explain the remarkable quiescence of most black holes in the local universe (Narayan et al. 1998).

6.4. Prospects

Our results illustrate how a variety of observational constraints can be brought to bear on the key questions of quasar and black hole evolution. In particular, we have extended the ideas of Soltan (1982) and Small & Blandford (1992) to show that incorporating the link between luminous accretion and black hole growth allows one to construct concrete physical models that are simultaneously constrained by multi-wavelength luminosity function measurements and estimates of black hole masses and accretion rates. For our models in §5, we chose a plausible but not unique set of initial conditions at $z = 2$ and evolved them forward in time under varying assumptions, always matching the observed optical QLF. The models then make distinguishable predictions for other observables.

With our current parameter choices, all of our models face some difficulty when confronted with recent estimates of X-ray luminosity functions and the local black hole mass function, as illustrated in Figures 10, 12, and 13. Models that fit the Ueda et al. (2003) hard X-ray QLF generally do not fit the Miyaji, Hasinger, & Schmidt (2001) soft X-ray QLF at the same redshift, and vice versa. A model incorporating luminosity and redshift dependence of the obscured quasar fraction might fare better, though some of the problem may still lie with the observational estimates themselves. Combining Tremaine et al.'s (2002) estimate of the $M - \sigma$ relation with Sheth et al.'s (2003) estimate of the distribution of galaxy velocity dispersions yields a mass function that lies well below our model predictions for $M > 10^9 M_\odot$, unless the intrinsic scatter of the $M - \sigma$ relation is ~ 0.5 dex, compared to Tremaine et al.'s estimate of ≤ 0.3 dex. Repairing this discrepancy would require either reducing the break mass at $z = 2$ substantially below $M_* = 10^9 M_\odot$ or dropping our assumed double power-law form of $n(M)$ and adopting a mass-dependent $p(\dot{m})$ to reproduce the Boyle et al. (2000) QLF. For the present, we do not want to draw strong conclusions from these discrepancies, since we have not thoroughly assessed the observational uncertainties, and we have not investigated the extent to which a failing model can be "fixed up" by adjusting its parameters (e.g., the initial shape of the black hole mass function), while retaining its essential features (e.g., a large fraction of obscured systems).

The $z = 0$ black hole mass function can be reasonably well estimated from current data, at least at masses $M \leq 10^9 M_\odot$ where the form and scatter of the $M - \sigma$ relation are well constrained, and it is a fundamental boundary constraint on any evolution model. For a comprehensive attempt to match observations, therefore, it probably makes sense to impose this constraint *a priori* on all models, and integrate the evolutionary equations *backward* in time. In our framework, this approach is just as easy as integrating forward, even if it is less intuitive. In effect, one takes the known black hole mass function today, infers the accretion rate distribution by matching the luminosity function, steps backward by removing the implied amount of mass from each bin of the mass function, and repeats. We will apply this approach to available observations in future work. Given the inevitable uncertainties in the observational data, radiative efficiencies, and bolometric corrections, there are likely to be some degeneracies in the solutions, but we can hope that models that differ in fundamental rather than incidental features will remain observationally distinguishable. Based on the results found here, we suspect that the primary source of uncertainty will be the mass dependence of $p(\dot{m}|M)$, which requires accurate measurements of both the QLF and the masses of active black holes to pin down empirically. Mergers look like the other most difficult problem, though in this case there are good theoretical ideas about what the merger rates of dark halos and galaxies should be (e.g., Taylor & Babul 2001), and these can be incorporated into model calculations.

The traditional picture of quasars as a population of supermassive black holes growing by accretion seems more secure than ever. Many open questions remain about the roles of black hole mass, accretion rate, radiative efficiency, and SED shape in determining quasar luminosities, about the properties of accretion flows at low and high accretion rates, about the importance of black hole mergers and obscured accretion as drivers of black hole growth, and about the relations among populations observed at different wavelengths. Our work highlights a number of areas where observational advances will be crucial to answering these questions. These include improved determination of the local black hole mass function, better understanding of the dependence of radiative efficiency and SED shape on accretion rate, measurements of the luminosity function at different wavelengths over the widest achievable range in luminosity and redshift, estimates of masses and accretion rates of active black holes as a function of redshift and luminosity, and indirect estimates of black hole space densities from host galaxy and quasar clustering studies. Fortunately, the observational situation is advancing rapidly, and many of these areas have seen substantial progress in the last few months alone, as discussed in §5. It is worth emphasizing the value of luminosity function determinations and black hole mass estimates that traverse the break in the QLF and extend as far below as possible. Accurate characterization of this regime is crucial for separating the roles of $n(M)$ and $p(\dot{m})$ in shaping the luminosity function, which in turn is necessary for understanding the contribution of sub-Eddington accretion rates to black hole mass evolution. These lower luminosities are also where optical and X-ray evolution appear to be radically different, and better measurements of the joint X-ray, optical, and IR luminosity functions are needed to pin down the origin of these differences. The emerging data on black hole and quasar evolution are complex, complementary, and rich. We hope that the physical modeling approach described in this paper will prove useful in exploiting their power.

We thank Jordi Miralda-Escudé for crucial suggestions during the early stages of this work. We also thank Jordi and Martin Haehnelt, Paul Martini, and Smita Mathur for providing detailed comments on the manuscript that led to significant improvements. We have also benefited from discussions with many other colleagues, including Ramesh Narayan, Patrick Osmer, Bradley Peterson, Richard Pogge, Hans-Walter Rix, and Marianne Vestergaard. The discussion of quasar hosts in §5.5 was stimulated by conversations with Hans-Walter Rix. This work was supported by NSF grants AST-0098515 and AST-0098584.

REFERENCES

Abramowicz, M. A., Czerny, B., Lasota, J. P., & Szuszkiewicz, E. 1988, ApJ, 332, 646

Aller, M. C. & Richstone, D. 2002, AJ, 124, 3035

Anderson, S.F., et al. 2003, AJ, in press, astro-ph/0305093

Barger, A. J., Cowie, L. L., Bautz, M. W., Brandt, W. N., Garmire, G. P., Hornschemeier, A. E., Ivison, R. J., & Owen, F. N. 2001, AJ, 122, 2177

Barger, A. J., Cowie, L. L., Brandt, W. N., Capak, P., Garmire, G. P., Hornschemeier, A. E., Steffen, A. T., & Wehner, E. H. 2002, AJ, 124, 1839

Begelman, M. C. 1978, MNRAS, 243, 610

Begelman, M. 2002, ApJ, 568, L97

Blandford, R. D., & Begelman, M. C. 1999, MNRAS, 303, 1

Brandt, W.N., et al. 2001, AJ, 122, 281

Boyle, B. J., Shanks, T., Croom, S. M., Smith, R. J., Miller, L., Loaring, N., & Heymans, C. 2000, MNRAS, 317, 1014

Cavaliere, A., & Vittorini, V. 2000, ApJ, 543, 599

Cavaliere, A., & Vittorini, V. 2002, ApJ, 570, 114

Chokshi, A., & Turner, E. L. 1992, MNRAS, 259, 421

Comastri, A., Fiore, F., Vignali, C., Matt, G., Perola, G.C., LaFranca, F. 2001, MNRAS, 327, 781

Comastri, A., Setti, G., Zamorani, G., & Hasinger, G. 1995, A&A, 296, 1

Corbett, E. A., et al 2003, MNRAS, 343,705

Cowie, L. L., Barger, A. J., Bautz, M. W., Brandt, W. N., & Garmire, G. P. 2003, ApJ, 584, 57

Croom, S. et al. 2003, to appear in AGN Physics with the Sloan Digital Sky Survey, ed. G.T. Richards & P.B. Hall (San Francisco:ASP), astro-ph/0310533

Czerny, B., Nikołajuk, M., Piasecki, M., & Kuraszkiewicz, J. 2001, MNRAS, 325, 865

Czerny, B., Nikolajuk, M., Rozanska, A., Dumont, A. M., Loska, Z., & Zycki, P.T. 2003, A&A, in press, astro-ph/0309242

Dunlop, J. S., McLure, R. J., Kukula, M. J., Baum, S. A., O'Dea, C. P., Hughes, D. H. 2003, MNRAS, 340, 1095

Elvis, M., Risaliti, G., & Zamorani, G. 2002, ApJ, 565, 75

Elvis, M., Wilkes, B. J., McDowell, J. C., Green, R. F., Bechtold, J., Willner, S. P., Oey, M. S., Polomski, E., & Cutri, R. 1994, ApJS, 95, 1

Fan, X., et al. 2001, AJ, 121, 54

Efstathiou, G., & Rees, M. J. 1988, MNRAS, 230, 5

Esin, A. A., McClintock, J. E., & Narayan, R. 1997, ApJ, 489, 865

Fabian, A. C. 2003, to appear in Carnegie Observatories Astrophysics Series, Vol. 1: Coevolution of Black Holes and Galaxies, ed. L. C. Ho (Cambridge: Cambridge Univ. Press), astro-ph/0304122.

Fabian, A. C., & Iwasawa, K. 1999, MNRAS, 303, 34

Fiore, F., La Franca, F., Giommi, P., Elvis, M., Matt, G., Comastri, A., Molendi, S., & Gioia, I. 1999, MNRAS, 306, 55

Fiore, F. et al. 2003, A&A, 409, 79

Gebhardt, K., et al. 2000, ApJ, 539, 13

Gebhardt, K., et al. 2000, ApJ, 543, 5

Giacconi, R., et al. 2002, ApJS, 139, 369

Gilli, R., Salvati, M., & Hasinger, G. 2001, A&A, 366, 407

Haenhelt, M. G., Natarajan, P., & Rees, M. J. 1998, MNRAS, 300, 817

Haenhelt, M. G., & Rees, M. J. 1993, MNRAS, 263, 168

Haiman, Z., & Loeb, A. 2001, ApJ, 552, 459

Haiman, Z., & Hui, L. 2001, ApJ, 547, 27

Hasinger, G. 2003, to appear in High Energy Processes and Phenomena in Astrophysics, IAU Symposium 214, eds. X. Li, Z. Wang, & V. Trimble, astro-ph/0301040

Ho, L. 2001, in IAU Colloq. 184, AGN Surveys, eds. R. F. Green, E. Ye. Khachikian, & D. B. Sanders (San Francisco: ASP), astro-ph/0110438

Ho, L. 1999, ApJ, 516, 672

Kaiser, N. 1984, ApJ, 284, 9

Katz, J. 1977, ApJ, 215, 265

Kauffmann, G., & Haehnelt, M. 2000 MNRAS, 311, 576

Kauffmann, G. & Haehnelt, M. G. 2002, MNRAS, 332, 529

Kauffmann, G., et al. 2003, MNRAS, submitted, astro-ph/0304239

Kuraszkiewicz, J. K., Wilkes, B. J., Czerny, B., Mathur, S., Brandt, W. N., & Vestergaard, M. 2000, New Astronomy Review, 44, 573

Laor, A. 1998, ApJ, 505, L83

Lee, M. H. 2000, Icarus, 143, 74

Lynden-Bell, D. 1969 Nature, 223, 690

Madau, P., Rees, M. J., Volonteri, M., Haardt, F., & Oh, S. P. 2003, ApJ, submitted, astro-ph/0310223

Martini, P., & Weinberg, D. H. 2001, ApJ, 547, 12

Martini, P. 2003, to appear in Carnegie Observatories Astrophysics Series, Vol. 1: Coevolution of Black Holes and Galaxies, ed. L. C. Ho (Cambridge: Cambridge Univ. Press), astro-ph/0304009

McLure, R. J., & Dunlop, J.S. 2002, MNRAS, 331, 795

McLure, R. J., & Dunlop, J.S. 2003, MNRAS, submitted, astro-ph/0310267

Menou, K., Haiman, Z., & Narayanan, V. K. 2001, ApJ, 558, 535

Merritt, D., & Ferrarese, L. 2000, MNRAS, 320, 30

Merritt, D. & Ferrarese, L. 2001, MNRAS, 320, L30

Miyaji, T., Hasinger, G., & Schmidt, M. 2001, A&A, 353, 25

Mo, H.J., & White S.D.M. 1996, MNRAS, 282, 1096

Murali, C., Katz, N., Hernquist, L., Weinberg, D. H., & Dave, R. 2002, ApJ, 571, 1

Murray, S. D., & Lin, Douglas N. C. 1996, ApJ, 467, 265

Narayan, R., Mahadevan, R., & Quataert, E. 1998, in Theory of Black Hole Accretion Disks, eds. Marek A. Abramowicz, Gunnlaugur Bjornsson, & James E. Pringle. (Cambridge: Cambridge University Press), p. 148, astro-ph/9803141

Onken, C.A., & Peterson, B.M. 2002, ApJ, 585, 121

Priddey, R. S., Isaak, K. G., McMahon, R. G., & Omont, A. 2003, MNRAS, 339, 1183

Pei, Y. C. 1995, ApJ, 438, 623

Quataert, E., di Matteo, T., Narayan, R., Ho, L. 1999, ApJ, 525, 89

Quataert, E., & Gruzinov, A. 2000, ApJ, 539, 809

Redmount, I. H., & Rees, M. J. 1989, Com Ap, 14, 165

Rees, M. J. 1978, The Observatory, 98, 210

Richstone, D., et al. 1998 Nature, 395, 14

Salpeter, E. E. 1964, ApJ, 140, 796

Salucci, P., Szuszkiewicz, E., Monaco, P., & Danese, L. 1999, MNRAS, 307, 637

Sazonov, S. Yu., Ostriker, J.P., & Sunyaev, R.A. 2003, MNRAS, submitted, astro-ph/0305233

Schmidt, M. 1968, ApJ, 151, 393

Schmidt, M., Schneider, D.P., & Gunn, J.E. 1995, AJ, 107, 1245

Schneider, D. P., et al. 2002, AJ, 123,567

Setti, G. & Woltjer, L. 1989, A&A, 224, L21

Silk, J., Rees, M. J. 1998, A&A, 331, 1

Sheth, R. K. 1998, MNRAS, 295, 869

Sheth, R., et al. 2003, ApJ, 594, 225

Silk, J. & Takahashi, T. 1979, ApJ, 229, 242

Silk, J. & White, S. D. 1978, ApJ, 223, L59

Small, T. A., & Blandford, R. D. 1992, MNRAS, 259, 725

Soltan, A. 1982, MNRAS, 200, 115

Steffen, A. T., Barger, A. J., Cowie, L. L., Mushotzky, R. F., & Yang, Y. 2003, ApJ, 596, 23

Steidel, C. C., Hunt, M. P., Shapley, A. E., Adelberger, K. L., Pettini, M., Dickinson, M., Giavalisco, M. 2002, ApJ, 576, 653

Taylor, J.E., & Babul, A. 2001, ApJ, 559, 716

Tremaine, S., et al. 2002, ApJ, 574, 740

Ueda, Y., Akiyama, M., Ohta, K., & Miyaji, T. 2003, ApJ, in press, astro-ph/0308140

Valtonen, M. J., Mikkola, S., Heinamaki, P., & Valtonen, H. 1994, ApJS, 95, 69

Vestergaard, M. 2004, ApJ, 600, in press, astro-ph/0309521

Wandel, A., Peterson, B. M., & Malkan, M.A. 1999, ApJ, 526, 579

Warren, S.J., Hewett, P.C., & Osmer, P.S. 1994, ApJ, 421, 412

White, R.L., et al. 2000, ApJS, 126, 133

Wilkes, B. J., Schmidt, G. D., Cutri, R. M., Ghosh, H., Hines, D. C., Nelson, B., & Smith, P. S. 2002, ApJ, 564, 65

Wisotzki, L. 2000, A&A, 353, 853

Wolf, C., Wisotzki, L., Borch, A., Dye, S., Kleinheinrich, & Meisenheimer, K. 2003, A&A, 408, 499

Woo, J-H., & Urry, C. M. 2002, ApJ, 579, 530

Wyithe, J. S. B., & Loeb, A. 2003, ApJ, 595, 614

Yu, Q. & Tremaine, S. 2002, MNRAS, 335, 965

This preprint was prepared with the AAS LaTeX macros v5.0.

A. Luminosity Function for a Double Power-Law $p(\dot{m})$

For the double power-law $p(\dot{m})$ and double power-law $n(M)$, the convolution integral (8) for the luminosity function must be broken into three different regimes to account for the different efficiencies of the accretion modes. The total QLF is the sum of the QLFs produced by each accretion mode, $\Phi(L)_{\rm Total} = \Phi(L)_{\rm SE} + \Phi(L)_{\rm TD} + \Phi(L)_{\rm ADAF}$. The calculation is analogous to that in §3.1, though more tedious, and we omit the details. The solution depends on whether $\dot{m}_*$ lies in the thin-disk, super-Eddington, or ADAF regimes. For the first case, which is the one usually relevant to our models, the results for the three accretion modes are

$$\Phi(L)_{\rm SE} = \begin{cases} \frac{n_* p_*}{l(b+1)\dot{m}_*^b}\left[10^{b+1}-1\right]\left(\frac{L}{lM_*}\right)^{\alpha} & L < lM_* \\ \frac{n_* p_*}{l(b+1)\dot{m}_*^b}\left[10^{b+1}-1\right]\left(\frac{L}{lM_*}\right)^{\beta} & L > lM_* \ , \end{cases} \tag{A1}$$

$$\Phi(L)_{\rm TD} = \begin{cases} \frac{n_* p_*}{l}\left[\frac{\dot{m}_*^{a-\alpha}-0.01^{a-\alpha}}{(a-\alpha)\dot{m}_*^a}+\frac{1-\dot{m}_*^{b-\alpha}}{(b-\alpha)\dot{m}_*^b}\right]\left(\frac{L}{lM_*}\right)^{\alpha} & L < 0.01lM_* \\ \frac{n_* p_*}{l}\left[\frac{(\beta-\alpha)\dot{m}_*^{-a}}{(a-\alpha)(a-\beta)}\left(\frac{L}{lM_*}\right)^{a}+\frac{(b-a)\dot{m}_*^{-\alpha}}{(a-\alpha)(b-\alpha)}\left(\frac{L}{lM_*}\right)^{\alpha}\right. & \\ \qquad \left.+\frac{1}{(b-\alpha)\dot{m}_*^b}\left(\frac{L}{lM_*}\right)^{\alpha}-\frac{0.01^{a-\beta}}{(a-\beta)\dot{m}_*^a}\left(\frac{L}{lM_*}\right)^{\beta}\right] & 0.01lM_* < L < \dot{m}_* lM_* \\ \frac{n_* p_*}{l}\left[\frac{(\beta-\alpha)\dot{m}_*^{-b}}{(b-\alpha)(b-\beta)}\left(\frac{L}{lM_*}\right)^{b}+\frac{(b-a)\dot{m}_*^{-\beta}}{(a-\beta)(b-\beta)}\left(\frac{L}{lM_*}\right)^{\beta}\right. & \\ \qquad \left.+\frac{1}{(b-\alpha)\dot{m}_*^b}\left(\frac{L}{lM_*}\right)^{\alpha}-\frac{0.01^{a-\beta}}{(a-\beta)\dot{m}_*^a}\left(\frac{L}{lM_*}\right)^{\beta}\right] & \dot{m}_* lM_* < L < lM_* \\ \frac{n_* p_*}{l}\left[\frac{\dot{m}_*^{a-\beta}-0.01^{a-\beta}}{(a-\alpha)\dot{m}_*^a}+\frac{1-\dot{m}_*^{b-\beta}}{(b-\alpha)\dot{m}_*^b}\right]\left(\frac{L}{lM_*}\right)^{\beta} & L > lM_* \ , \end{cases} \tag{A2}$$

and

$$\Phi(L)_{\rm ADAF} = \begin{cases} \frac{n_* p_* 0.01^{\alpha+1}}{l(a-2\alpha-1)\dot{m}_*^a}\left[0.01^{a-2\alpha-1}-10^{-4(a-2\alpha-1)}\right]\left(\frac{L}{lM_*}\right)^{\alpha} & L < 10^{-4}lM_* \\ \frac{n_* p_*}{l\dot{m}_*^a}\left[\frac{2(\beta-\alpha)}{(a-2\beta-1)(a-2\alpha-1)}\left(\frac{L}{lM_*}\right)^{\frac{a-1}{2}}\right. & \\ \qquad \left.-\frac{10^{-4(a-2\beta-1)}}{a-2\beta-1}\left(\frac{L}{lM_*}\right)^{\beta}+\frac{0.01^{a-2\alpha-1}}{a-2\alpha-1}\left(\frac{L}{lM_*}\right)^{\alpha}\right] & 10^{-4}lM_* < L < 0.01lM_* \\ \frac{n_* p_* 0.01^{\beta+1}}{l(a-2\beta-1)\dot{m}_*^a}\left[0.01^{a-2\beta-1}-10^{-4(a-2\beta-1)}\right]\left(\frac{L}{lM_*}\right)^{\beta} & L > 0.01lM_* \ , \end{cases} \tag{A3}$$

where the values $\dot{m}_{\rm crit} = 0.01$, $\dot{m}_{\rm max} = 10$, and $\dot{m}_{\rm min} = 10^{-4}$ are explicitly included in the solutions. The three regimes differ because of the different dependence of $\epsilon_{0.1}$ on $\dot{m}$ and because $\dot{m}_*$ lies in the thin-disk regime, breaking that integral into more parts. Note that the slopes at the high and low luminosity end of each mode are equal to the mass function slopes (α and β), for the reasons discussed in §3.1.

B. Critical Slopes

We gain some insight into results for general $p(\dot{m})$ and $n(M)$ by considering cases in which $p(\dot{m}) = \dot{m}^a$ is a pure power-law in some range $\dot{m}_{\rm min} - \dot{m}_{\rm max}$ and $n(M) = M^\alpha$ is a pure power-law in some range $M_{\rm min} - M_{\rm max}$. Analysis of such cases reveals a number of critical slopes where the character of the solutions changes. For the critical $p(\dot{m})$ slope $a = -2$, each logarithmic range of $\dot{m}$ contributes equally to black hole growth and, if $\epsilon_{0.1}$ is constant, to emissivity of the quasar population. When $a \ll -2$, growth and emissivity are dominated by low accretion rates (near $\dot{m}_{\rm min}$), and when $a \gg -2$ high accretion rates (near $\dot{m}_{\rm max}$) dominate. Our double power-law models have a low-$\dot{m}$ slope $a > -2$ and a high-$\dot{m}$ slope $b < -2$, so that the largest contributions to growth and emissivity are from accretion rates near $\dot{m}_*$, which we usually take (at least at high redshift) to lie in the thin-disk range $0.01 < \dot{m}_* < 1$. These parameter choices make our results relatively insensitive to our assumptions about efficiencies and the form of $p(\dot{m})$ in the ADAF and super-Eddington regimes. However, low-$\dot{m}$ or high-$\dot{m}$ accretion may be more important in the real universe, at least at some redshifts, in which case the form of $p(\dot{m})$ and behavior of the accretion physics in these regimes would have a larger impact on observable properties of the quasar population.

The distribution of active black hole masses at a fixed luminosity depends on both the $p(\dot{m})$ and $n(M)$ slopes. When $\alpha < a+1$ and $\epsilon_{0.1} = \text{constant}$, the black hole mass function is steep enough that low mass black holes with high accretion rates predominate (eq. 21) — i.e., the most common black holes at a given luminosity L are either those with $M = M_{\rm min}$ or those with the maximum accretion rate and $M = L/(\epsilon_{0.1} l \dot{m}_{\rm max})$. Conversely, when $\alpha > a+1$, high mass black holes with low accretion rates predominate. For more general forms of $n(M)$ and $p(\dot{m})$, the roles of "minimum" and "maximum" values are in practice played by values where the slope of the distribution changes or where there is a break in efficiency. For example, if $\alpha < a+1$ and the power-law behavior of $p(\dot{m})$ extends to $\dot{m} \approx 1$, then decreasing efficiency in the super-Eddington regime comes into play, and the luminosity function is dominated by black holes radiating near the Eddington luminosity. This is the typical behavior for our double power-law models at high redshift, where $\dot{m}_*$ is close to unity and the low-$\dot{m}$ slope of $p(\dot{m})$ and high mass slope of $n(M)$ easily satisfy $\beta < a+1$. At low redshift we have low values of $\dot{m}_*$, so the high-$\dot{m}$ slope of $p(\dot{m})$ becomes important. Even here we usually have $\beta < b+1$, so that systems with $\dot{m} \approx \dot{m}_*$ dominate the high luminosity end of $\Phi(L)$, but because the slopes that we adopt ($b = -3$, $\beta = -3.4$) are not so far from the critical relation, systems at a given luminosity span a wide range of black hole masses and accretion rates.

Two more critical slopes arise when we ask whether the space density of black holes of mass M increases or decreases with time. If accretion drives the evolution of $n(M)$, then the critical slope is -1: for a mass function steeper than $n(M) \propto M^{-1}$, accretion increases $n(M)$, while for a shallower slope the number of black holes "lost" to higher masses exceeds the number gained from lower masses, driving $n(M)$ down with time. For merger driven growth, the corresponding critical slope is -2, at least if objects merge with others of equal mass as in the simple models considered here. The difference in slopes arises because accretion adds mass to the black hole population while mergers do not, and the critical slope for mergers is steeper still if black hole ejection or gravitational radiation reduce the mass of the surviving merger product below the combined mass of its progenitors. For black hole mass functions whose asymptotic slopes match the asymptotic slopes of the Boyle et al. (2000) luminosity function, mergers drive $n(M)$ up in the high mass regime and down in the low mass regime. Accretion increases $n(M)$ in both regimes, but the increase is slow for low masses and rapid for high masses.